South African Radio League

Radio Amateur Examination Study Guide

REVISED NOVEMBER 2007

WE ACKNOWLEDGE WITH THANKS THE ORIGINAL COMPILATION OF THIS STUDY GUIDE BY ANDREW ROOS IN 2005

Table of Contents

Chapter 1 - Introduction to Amateur Radio

Amateur radio is a hobby that involves experimenting with radio (and related technologies like television or radar) for fun and education. It is also known as "Ham Radio" and radio amateurs are sometimes referred to as "hams". Like most hobbies, there are many different activities that fall under its umbrella.

Communicating with other Radio Amateurs

Using radio to communicate with other amateurs is one of the foundations of the hobby. Most amateurs have a radio station of their own, which can range from a simple single-band handheld transceiver (a combination of a *transmitter* and a *receiver* is known as a *transceiver*) for talking to others in the same town, to a sophisticated station that is capable of worldwide communication. Many clubs also have club stations that are available for use by club members.

Radio amateurs communicate in many different *modes*. The most common are by voice (known as *phone* although it does not use the telephone system), Morse code (also referred to a *CW*) and various digital modes including *slow-scan television*. The contents of an amateur communication (known as a *QSO*) range from the briefest exchange of name and location, up to long conversations (known as *rag-chews*) that may last an hour or more.

Amateur radio is not like the phone system since you generally can't dial a particular station. If you want to speak to a particular person, then you must agree a time and a frequency where you will meet – this is known as a schedule, or "sked" for short. Otherwise you can just speak to whoever happens to be listening and is interested in a chat, which is a great way to make new friends. There are also some regularly scheduled networks (or "nets") where operators who share a common interest get together at a particular time and frequency to exchange ideas.

Collecting QSL Cards

After communicating with another amateur (especially one in a foreign country) it is customary to send a QSL card, which is a postcard-sized card with information about yourself and your station, and details of the QSO such as the date, time, frequency, mode and the callsign of the station worked. Many amateurs take a great deal of pride in their QSL cards, which are works of art. As well as being something to display and a nice reminder of the contact, QSL cards are often required if you wish to claim a contact for an award (see below).

Building Radio and Electronics Equipment

Many amateurs build at least some of their equipment. Some build equipment from purchased kits or from plans found in amateur radio magazines. Others build their equipment from scratch, doing all the necessary design and sourcing the components themselves. The complexity ranges from simple projects, such as a computer soundcard interface that can be built in an evening to complete radio transmitters and receivers that may take months or years of work. Today microprocessors and digital signal processing (DSP) is an increasingly important part of the hobby, so building equipment may also involve writing the necessary micro-controller or DSP programs. Of course if you do not enjoy electronics, then everything you need to participate in the hobby can be purchased off the shelf.

Building Antennas

Most amateurs build at least some of their own antennas. Antennas may range from a simple wire antenna suspended from a tree, to a complex multi-element beam sitting on top of a large tower. Antenna projects can be very rewarding as good results may be obtained from fairly simple designs. There are a number of software packages available that allow you to design an antenna and model its performance before you invest in the construction of the antenna.

Public Service and Emergency Communications

Radio amateurs have a proud history of making their skills and equipment available for public service and emergency communications. On the public service side, amateurs provide communications for many sporting events such as rallies, marathons and cycle tours where their ability to communicate effectively from remote places is of great assistance to the organizers.

Many amateurs also ensure that their radio stations have some alternative power source (which could be batteries, a generator, or solar power) so that they can continue to provide communications in the event that a natural disaster disrupts the telephone and power distribution systems. In South Africa, Hamnet, a special interest group of the South African Radio League, coordinates amateur emergency communications.

DX'ing

"DX'ing" means communicating with as many different places as possible, often in order to qualify for certificates and awards. (The term comes from the use of "DX" as an abbreviation for "long distance".) There are many different awards, including:

- The DXCC (DX Century Club) certificate, which you qualify for by communicating with 100 or more different countries.
- The Worked All ZS award, for contacting 100 stations in the various regions of South Africa (the award's name comes from the fact that "ZS" is one of the callsign prefixes assigned to South African radio stations).
- The Islands on the Air (IOTA) awards, which are given for communicating with stations located on islands.
- The Summits on the Air (SOTA) awards for communicating with mountaintop stations.

DXpeditions

Because DXers are always on the lookout for countries, islands, mountains or provinces that they have not worked before, there is often a flurry of interest and activity when a rare country or island is "activated" by some intrepid radio amateur setting up a station. Expeditions to unusual places for the purpose of setting up and operating a radio station there are called "DXpeditions", and participating in DXpeditions is itself a very rewarding and challenging activity.

Contests

Contests bring out the competitive nature of some radio amateurs, who enjoy the challenge to contact as many different stations as possible over a predetermined period of anything from an hour or two up to 48 hours. Contests may be run on a local, national, regional or international basis and may attract anything from 10 to 5 000 contestants. Many contests have several entry categories to allow similarly equipped stations to compete amongst each other.

Satellite Communications

The amateur community has successfully launched a number of small communications satellites for the use of radio amateurs around the world. Communicating with other amateurs via satellite (or via the earth's natural satellite, the moon) gives radio amateurs an unparalleled opportunity to learn about the technology that underlies much of the modern era of communications. Because amateurs themselves develop these satellites as a cooperative, non-profit venture, those who are interested in the design and construction of satellites also have the opportunity to study the designs and may eventually be able to contribute to new amateur satellite projects.

Maritime and Off-Road Communications

The maritime and off-road communities are increasingly turning to amateur radio for their communication needs. Thousands of small craft such as yachts make use of the services provided by maritime nets which pass on weather reports and crucial safety information, allow mariners to access email and assist in the search for missing boats. Off-roaders who venture into uninhabited areas can also benefit from amateur communications, both between vehicles within a party and also back to a "home base" or to summon assistance in an emergency.

License Requirements in South Africa

In order to operate an amateur radio station you must have a license issued by the Independent Communications Authority of South Africa, ICASA. When you are issued with a license you will also be given a unique callsign. Every amateur has a callsign, which is used to identify him or her on the air. South African amateur callsigns consist of the letters "ZU", "ZR" or "ZS" followed by a single digit indicating the region of the country in which you are located, followed by one to three letters. Using the callsign "ZS1AN" as an example, the "ZS" indicates that it is an Unrestricted license, the "1" shows that the holder resides in the Western Cape, and the letters "AN" distinguishes the holder from all the other holders of an unrestricted license in the Western Cape. Every time you make a transmission from an amateur radio station you are required to identify yourself using your callsign. There are three different classes of license:

The Entry Level (ZU) License

The Entry level license is a simple entry point into the hobby. It has restricted privileges in the High Frequency (HF) and Very High Frequency (VHF) bands, with a maximum power output of 20 W for single sideband (voice) transmissions. In order to obtain a Novice license you must pass a (Class B) Radio Amateurs' Examination

The Restricted (ZR) License

The Restricted license has full privileges on the Very High Frequency (VHF) and Ultra High Frequency (UHF) bands, which are typically used for short-range communication and for communication via satellite. ZR licence holders have restricted access to HF frequencies with a power limit of 20 dBW

To obtain a Restricted (ZR) license you must pass the full (Class A) Radio Amateur's Examination. This is the study material for the Class A examination, so if you pass the examination at the end of the course you will be entitled to a Restricted license. Restricted licenses have callsigns beginning with "ZR".

The Unrestricted (ZS) License

The Unrestricted license has full privileges on all the bands allocated for use by radio amateurs. In most cases the maximum power output is 400W for single sideband (voice) transmissions. To obtain an Unrestricted license you must pass the full (Class A) Radio Amateurs' Examination, as well as furnish proof of your ability to correctly set up, adjust and operate an amateur HF transceiver. In addition you have to complete an assessment prescribed by the SARL, the National Body for Amateur Radio, demonstrating advanced knowledge of theoretical or practical aspects of amateur radio. (Visit www.sarl.org.za for the details)

The Radio Amateurs' Examination

The Radio Amateur's Examination is held twice each year, in May and October. It consists of two papers: *Regulations and Operating Procedures* and *Technical*. The *Regulations and Operating Procedures* paper has 30 multiple-choice questions and the *Technical* paper consists of 60 multiple-choice questions.

The examination is set and administered by the South African Radio League, the National Body for Amateur Radio in South Africa. ICASA is the Independent Communications Authority of South Africa, a statutory body that regulates the communications industry. The examination fee changes from time to time, so ask your course instructor what the current fee is, or consult the website of the South African Radio League, www.sarl.org.za

Restrictions on the Use of an Amateur Radio Station

The Radio Regulations include some restrictions on the use of an amateur radio station. It is important that you understand these in case you find that what you had planned to do with your amateur radio license is not permitted!

1. Amateur radio stations may not be used for broadcasting. Amateur radio is intended for direct "one-on-one" communications with other amateurs, and not as a community broadcasting service.

2. Amateur radio stations may only transmit music under very specific conditions, which are intended to ensure that they do not become pirate broadcast stations.

3. No products or services may be advertised on amateur radio.

4. Amateur radio stations may not transmit messages for reward.

5. Amateur radio stations may not be used to transmit business messages that could be sent using the public telecommunications service.

6. Amateur radio stations may not be used to transmit indecent, offensive, obscene, threatening or racist comments.

7. Amateur radio stations may not be used to pass third-party traffic (in other words, messages that originate from anyone other than the amateur who is operating the station) except during an emergency.

This chapter was intended to give you an idea of what amateur radio is all about, what the license requirements are, and what legal restrictions there are on what can be transmitted by amateur radio stations. We hope this will have helped you to decide that amateur radio is a hobby that you wish to participate in. We would like to take this opportunity to welcome you to the amateur community and hope that you will find this course interesting and worthwhile.

Chapter 2 - Basic Electrical Concepts

Atoms and Electrons

The matter that we interact with every day – solid objects like desks and computers, liquids like water and gasses like the air we breathe – consists of atoms. Atoms are tiny and invisible to the naked eye and were once thought to be the ultimate indivisible constituents of matter, but we now know that they are themselves made up of various sub-atomic particles. For the purposes of this discussion, atoms can be thought of as a very small central nucleus that is surrounded by a cloud of electrons. Electrons are not simple particles like miniature planets surrounding the nucleus, but are "smeared out" in space so that even a single electron can form a cloud around a nucleus.

The nucleus consists of one or more protons, which are positively charged particles, usually accompanied by some neutrons, which are uncharged (electrically neutral) particles. So the overall charge of a nucleus is always positive, from the positively charged protons. Electrons are negatively charged, and since opposite charges attract, the negatively charged electrons are attracted to the positively charged nucleus, which is what makes the electrons stay close to the nucleus.

Point to remember: *Unlike charges attract, like charges repel.*

Of course, since like charges repel, you might ask what stops the positively charged protons in the nucleus from bursting apart and destroying the atom. The answer is that another force called the "strong nuclear force" holds the nucleus together. The strong nuclear force is stronger at the very short distances characteristic of an atomic nucleus than the repulsive electromagnetic force between the positively charged protons.

Visible amounts of matter contain huge numbers of atoms. For example, a copper cube 1mm on each side would weigh less than one hundredth of a gram, but would contain about 85 000 000 000 000 000 000 atoms!

Conductors and Insulators

In some materials, such as copper, some of the electrons are not very strongly bound to their nuclei. These electrons are free to move around in the material, as long as other electrons replace them when they move. If they were not replaced then the area they left would have more protons than electrons, giving it an overall positive charge. This would attract electrons back there and make it harder for other electrons to leave, since the negatively charged electrons would be attracted by the positive overall charge.

Materials in which some of the electrons can move around relatively freely conduct electricity and are known as "conductors". Materials in which all the electrons are tightly bound to their nuclei and cannot move around do not conduct electricity and are known as "insulators".

Most metals are conductors. These include silver (which is the best conductor of all, but too expensive for most uses), copper (a very good conductor at a more reasonable price), aluminium (which is ideal for weight-sensitive applications like overhead cables), mercury (for when you need a good conductor that is a liquid at room temperature) and solder (an alloy, often of tin and lead, with a low melting point that is used to connect electrical components together).

Good insulators include most plastics, glass, plexiglass, mica, rubber and dry wood. (Note that water is not an insulator, so anything wet is likely to conduct electricity, especially if you did not intend it to.)

Electric Current

When we speak of a material conducting electricity, we mean that electric currents can flow through that material. But what is an electric current?

Definition: *An electric current is a flow of charge.*

Any time that charge is flowing – that is, moving in a relatively consistent direction – there is an electric current. Since charge is normally associated with particles of one sort or another, a flow of charge usually entails a flow of charged particles, such as electrons. The particles that carry the charge are known as "charge carriers".

The size of an electric current is expressed in amperes, named after the French physicist André-Marie Ampère (1775-1836) who was one of the pioneers in studying electricity. The official abbreviation is "A", but unofficially amperes are often referred to as "amps".

When an electric current flows through ordinary conductors like copper, the charge carriers are electrons, so the flow of electric current corresponds to a flow of electrons. However because electrons are negatively charged, electrons flowing from left to right through a wire would constitute a *negative* current flowing from left to right in the wire. This is usually expressed as a *positive* current flowing in the opposite direction, from right to left in this case. So the electric current is generally considered to flow in the *opposite* direction from the electrons that carry it!

In order to emphasise this distinction, one may talk about the *conventional current* that flows in the opposite direction from the flow of electrons. However whenever someone refers to just an "electric current" you should assume that they are talking about a conventional current, so if the charge carriers are negatively charged particles like electrons then the direction in which the current flows will be the opposite direction to the flow of charge carriers.

You can imagine a (conventional) electric current flowing in a wire to be similar to water flowing though a pipe. Here the magnitude (size) of the current would correspond to the volume of water flowing through the pipe each second.

Electric Potential

Having established that an electric current is a flow of charge, the next question is: what makes the charge flow? The answer is electric potential. Since unlike charges attract, if you apply a positive potential to one side of a copper wire and a negative potential to the other side, loosely bound electrons in the rod will be attracted towards the positive potential and repelled by the negative potential, causing electrons to move from the negative side to the positive side. In other words, a conventional current will flow from the positive side of the wire to the negative side. Electric potential is always measured between two points.

Definition: *The electric potential between two points is the amount of energy that it would take to move one unit of charge from the point of lower potential to the point of higher potential.*

Since energy is measured in Joules and charge in Coulombs, the unit of electric potential is Joules per Coulomb. This unit is named the "volt" with the abbreviation "V", named after the Italian scientist Count Alessandro Volta (1745-1827) who invented the battery. Electric potential is commonly referred to simply as "voltage".

Electric potential can be thought of as the pressure that a power source like a pump creates in the water it pumps through a pipe. The higher the pressure (voltage), the greater the quantity of water flowing through the pipe per second (current).

Units and Abbreviations

If you measure or calculate the amount of something, you usually need to specify the unit of measurement. For example, if you weigh something and then say that it weighs "10", this does not mean very much unless you specify the unit of measurement – 10 grams, or 10 kilograms, or 10 milligrams.

The units of measure used in this course are the standard S.I. units that are used throughout most of the Western world. Each unit has a name, like "volt" or "ampere", and a corresponding abbreviation, like "V" for volt and "A" for ampere. This saves time when writing quantities – for example a current of "10 A" rather than "10 amperes".

There are also a number of standard prefixes, which are used to indicate quantities a thousand or a million or more times bigger or smaller than the basic unit. For example, the prefix "milli" which is abbreviated "m" means "one thousandth of", so one milligram – written as "1 mg" – means one thousandth of a gram. The following prefixes are widely used in electronics:

Name	Abbreviation	Scale Factor	Scientific Notation
pico	p	÷ 1 000 000 000 000	10^{-12}
nano	n	÷ 1 000 000 000	10^{-9}
micro	μ	÷ 1 000 000	10^{-6}
milli	m	÷ 1 000	10^{-3}
kilo	k	* 1 000	10^{3}
mega	M	* 1 000 000	10^{6}
giga	G	* 1 000 000 000	10^{9}

Scientific Notation

The column headed "scientific notation" may not be familiar to you. Because scientists work with very small and very large numbers, it would be inconvenient for them to have to keep writing many zeroes after the large numbers, or a decimal point and many zeros before the small numbers. So they use the fact that multiplying by ten to the power of any positive number effectively adds that many zeros at the end of the number. So for example the speed of light is about $3 * 10^8$ m/s which means "3 followed by 8 zeros", or 300 000 000 m/s.

Another way of thinking of this is that it is equivalent to moving the decimal point 8 places to the right, and introducing as many zeros as are necessary to do so. This is helpful when the number already has a decimal point, for example "$2{,}998 * 10^8$". Then you can't simply thing of adding zeros, since adding 8 zeros to "2,998" would give you "2,998 000 000 00" which represents the same number (although to a greater precision). However if you instead think about moving the decimal point 8 places to the right and adding zeros as necessary you get the correct result, which is 299 800 000. The power of ten – in this case, 8 – is known as the "exponent" and most scientific calculators have a key marked "E" or "Exp" which is used to enter numbers in this format.

Similarly, a negative exponent means you move the decimal point that number of places to the *left*, again filling in zeros as required. So for example, $1{,}6 * 10^{-19}$ is equivalent to 0,000 000 000 000 000 000 16, a very small number indeed. (If you were wondering, it is the charge on an electron, in Coulombs.)

Summary

This module has introduced the concepts of electric charge, electric current, and electric potential. You have seen how the atomic structure of materials allows electric currents to flow through some materials, which we call conductors, but not through others, which we call

insulators. You have learnt the meaning of the prefixes that are used to scale units by powers of ten, and to understand numbers written in scientific notation.

Revision Questions

1. One of the following is not an electrical conductor:
a. Silver.
b. Aluminium.
c. Copper.
d. Mica.

2. One of the following is not an electrical insulator:
a. Mica.
b. Ceramic.
c. Plastic.
d. Copper.

3. The unit of electrical potential is the:
a. Ampere.
b. Amp.
c. Voltaire.
d. Volt.

4. A current of 15 μA is equivalent to:
a. $1.5 * 10^{-5}$ A.
b. $15 * 10^{-5}$ A.
c. $1.5 * 10^{6}$ A.
d. $15 * 10^{6}$ A.

5. A voltage of 20 000 V could be expressed as:
a. 20 μV.
b. 20 mV.
c. 20 kV.
d. 20 MV.

6. The charge carriers in solid copper that allow it to conduct electricity are:
a. Positively charged copper ions.
b. Negatively charged copper ions.
c. Positively charged electrons.
d. Negatively charged electrons.

7. Conventional current flows:
a. In the same direction as electrons are moving.
b. In the opposite direction to the flow of electrons.
c. At right angles to the flow of electrons.
d. From negative to positive.

8. An electric current always consists of:
a. A flow of electrons.
b. A flow of neutrons.
c. A flow of protons.
d. A flow of charge.

Chapter 3 - Resistance and Ohm's Law

In the last module we learnt that an electric current is a flow of charge that is caused by a potential difference between two points. We also saw that the greater the electric potential between two points joined by a conductor, the greater the current that would flow through the conductor.

However the electric potential between two points is not the only factor that determines the size of the current flowing between them. The current flow is also affected by a quality of the conductor, known as its resistance. The resistance of a conductor can be thought of as being the extent to which it resists the flow of current. The greater the resistance of a conductor, the lower the current that will flow through it for a given potential difference. Conversely the lower the resistance of the conductor, the greater the current that will flow through it for a given potential difference.

The German physicist Georg Ohm (1789-1854) discovered that the potential difference across a conductor is proportional the current that flows through the conductor. In other words if the current flowing through a conductor doubles then the voltage across that conductor will double. Conversely if the current through the conductor halves then the voltage across the conductor will also be halved.

Mathematically, this can be expressed by saying that the voltage across the conductor is equal to the current through the conductor multiplied by some constant (for that particular conductor). Ohm called this quantity the "resistance" of the conductor,

voltage = current * resistance

This relationship is known as "Ohm's Law".

The unit of resistance is the ohm. The abbreviation for the ohm is the Greek letter omega, which is written as Ω. A conductor has a resistance of one ohm if the application of a potential difference of one volt across the conductor causes a current of one ampere to flow through the conductor.

Resistance may be thought of as the opposition to the flow of electric current through a conductor or electric circuit.

Symbols in Mathematical Equations

In order to save time when writing out equations, it is common practice to use symbols to represent quantities rather that writing out the full names of quantities like "voltage" and "resistance" every time.

Certain symbols are commonly used to represent particular quantities. For example, "V" is commonly used to represent an electric potential (voltage), and "R" is usually used to represent a resistance. A current is usually not represented by "C" (which had already been assigned another meaning in physics, the speed of light) but instead by "I". Using these symbols instead of the full names of the quantities, Ohm's law us usually written as:

Ohm's Law: $V = I\,R$

Note that the multiplication sign between "I" and "R" is also omitted. In mathematics when two symbols are written next to each other it is assumed that they are to be multiplied together.

This form of Ohm's law is convenient if you know the current flowing through a conductor and the resistance of the conductor, and want to calculate the electrical potential (voltage) across the conductor. It shows that you can calculate the voltage by multiplying the current by the resistance.

For example, if a current of 5 A is flowing through a conductor with a resistance of 2 Ω then the electric potential (voltage) across the conductor can be calculated by replacing the "I" with 5 and the "R" with 2 in the equation for Ohm's law, giving

$$V = 5 * 2$$
$$= 10 \text{ V}$$

Note the somewhat confusing use of "V" both as the symbol for electric potential (voltage) and also as the abbreviation for the unit "volt". In this equation, the V on the left hand side (before the equals sign) is the symbol for electric potential. The V after the number 10 is the abbreviation for the unit, volts. The two meanings are not the same and you should take care not to confuse them. You should be able to work out the correct meaning from the context in which the "V" appears.

The symbol E is also used for electric potential. So you may see Ohm's law written as $E = I R$ instead of $V = I R$.

Rearranging Ohm's Law

This is all well and good if you know the current and the resistance and want to calculate the voltage. However Ohm's law can also be used to find either the current or the resistance if both the other quantities are known. This is done by using simple algebra to rearrange Ohm's law as follows:

$$V = I R$$

By dividing both sides by I you get

$$V/I = R$$
and $$R = V/I$$

This can be used to calculate the resistance of a conductor given the electric potential (voltage) across the conductor and the current flowing through it. Similarly, if you divide both sides of the original equation by R you get

$$V/R = I$$
and $$I = V/R$$

In this form, Ohm's law can be used to calculate the current flowing through a conductor given the electrical potential (voltage) across the conductor and the resistance of the conductor. You need to be able to use any of these three forms of Ohm's law in the examination.

Summary

Ohm's law states that the electric potential across a conductor is proportional to the current flowing through the conductor. It can be written as $V = I R$, where R is a constant of proportionality that is known the *resistance* of the conductor. Resistance may be thought of as the opposition to the flow of electric current through a conductor or electric circuit. Resistance is measured in *ohms*, with the abbreviation Ω. Ohm's law can be used to find the

electric potential across a conductor, or current flowing through the conductor, or the resistance of the conductor provided that the other two quantities are known.

Revision Questions

1 The opposition to the flow of current in a circuit is called:

a. Resistance.
b. Inductance.
c. Emission.
d. Capacitance.

2 The current through a 100 Ω resistor is 120 mA. What is the potential difference across the resistor?

a. 120 V.
b. 8,33 V.
c. 83,33 V.
d. 12 V.

3 The resistance value of 1 200 Ω can be expressed as:

a. 12 kΩ.
b. 1,2 kΩ.
c. 1,2 MΩ.
d. 0,12 MΩ.

4 How can the current be calculated when the voltage and resistance in a dc circuit is known?

a. I = E / R.
b. P = I * E.
c. I = R * E.
d. I = E * R.

5 A 12 V battery supplies a current of 0,25 A to a load. What is the input resistance of this load?

a. 0,02 Ω.
b. 3 Ω.
c. 48 Ω.
d. 480 Ω.

6 If 120 V is measured across a 470 Ω resistor, approximately how much current is flowing through this resistor?

a. 56,40 A.
b. 5,64 A.
c. 3,92 A.
d. 0,25 A.

7. How can the voltage across a resistor be calculated when the resistance of and current flowing through the resistor are known?

a. V = I / R.
b. V = R / I.
c. V = I R.
d. $V = I^2 R$.

8. The law that relates the current flowing through a conductor to the electric potential applied across the conductor is known as:

a. Kirchoff's Current Law.
b. Kirchoff's Voltage Law.
c. Kirchoff's Current and Voltage Law.
d. Ohm's Law.

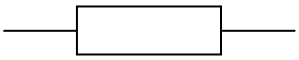
Chapter 4 - The Resistor and Potentiometer

Electronic circuits are usually constructed from components that can be purchased at electronics outlets. One such component is the *resistor*, which is simply a conductor that has a known resistance. Resistors are available in values ranging from a fraction of an ohm to several hundred mega-ohms.

Resistors also come in different tolerances. The tolerance shows how close the actual value of the resistor is guaranteed to be to its nominal value. For example, the actual resistance of a 1 kΩ resistor with a tolerance of 5% could range from 950 Ω (1 kΩ - 5%) to 1 050 Ω (1 kΩ + 5%).

Resistors also come in various power ratings. As you will see in a couple of modules time, the power dissipated by a resistance depends on the current flowing through the resistance and the voltage across the resistance. In order to cater for different requirements, resistors are usually available in power ratings from one eighth of a watt (0,125 W) to 5 W or more.

All electric components have symbols that can be used to draw diagrams showing how the components should be connected to create a particular circuit. These diagrams are known as "circuit diagrams" and the symbol for a resistor in a circuit diagram is:



In circuit diagrams, a plain line is used to represent a connection between two or more components, so the lines coming out of the left and right of the resistor represent its connections to the rest of the circuit. The resistor itself is the rectangle between these lines. In older circuit diagrams you may also see a resistor represented as a zigzag line, but we will not use that symbol. This symbol represents a simple fixed resistance. It has two connections (represented by the lines at the left and right) and there is a known resistance between these connections.

Different Types of Resistor

Resistors come in several different types, which are suited to specific applications:

- Carbon Film resistors are the most common, inexpensive, general-purpose resistors. They typically have a tolerance of ±5% and power ratings from 0,125 W to 2 W.

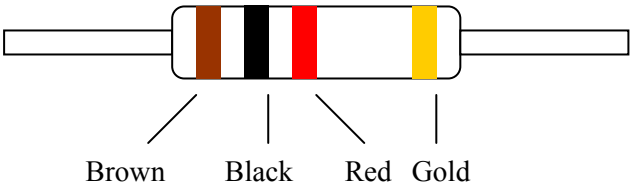
- Metal film resistors are often used when tighter tolerance is required (i.e. the resistor is guaranteed to be closer to the nominal value). Metal film resistors typically have tolerances of ±1% or better and power ratings from 0,125 W to 0,5 W.

- Wire wound resistors are used in D.C. applications when high power ratings are required. They are available in tolerances of ±5% or ±10% with power ratings from 2,5 W to 20 W or more. *Note that wire wound resistors should never be used in radio-frequency applications because they have unacceptably high inductance.* (You will learn about inductance in a future module).

- Resistor networks consisting of a number of resistors in various circuit configurations are supplied in packages that look like integrated circuits. They are intended for low-power applications and are especially useful when you need many resistors of the same value.

The Resistor Colour Code

Resistors are very small components, often only a few millimeters long, so if the value of a resistor (its nominal resistance, in ohms) were printed on the resistor it would be very difficult to read. So instead of printing the value onto resistors, a standard colour code is used where the value of the resistance is represented by three coloured bands, and the tolerance of the resistor by a fourth band. The following diagram represents not the circuit symbol for a resistor, but rather the physical resistor itself, showing the location of the colour bands.



From left to right the first two bands represent the first two digits in the value of the resistor. In this case, brown represents "1" and black represents "0" so the first two digits of the value are "10". The third colour band – red in this case – represents the number of zeros that should be added after the first two digits in the value (in other words, the exponent in scientific notation). Since red represents the value "2", two zeros must be appended to the first two digits, giving a value of 1000 Ω or 1 kΩ.

The last band, the gold one at the far right hand side, gives the tolerance of the resistor. Since gold means ±5%, the actual value of the resistor may range from 5% below the nominal value of 1 kΩ to 5% above the nominal value.

Colour	Digit	Multiplier	Tolerance
Black	0	* 1	
Brown	1	* 10	1%
Red	2	* 100	2%
Orange	3	* 1 000	
Yellow	4	* 10 000	
Green	5	* 100 000	
Blue	6	* 1 000 000	
Violet	7	* 10 000 000	
Grey	8	* 100 000 000	
White	9	* 1 000 000 000	
Gold			5%
Silver			10%

For each colour the table shows you the digit that it represents when it occurs in the first two bands, the multiplier it represents when it appears in the third band, and the tolerance that it represents when it occurs in the last band.

Resistors with tight tolerances, such as 1% or 2% resistors, may have an extra band in the colour code. In this case, the first *three* bands represent the first three digits of the value so that the value of the resistor can be represented more precisely. The remaining bands represent the multiplier and tolerance as before.

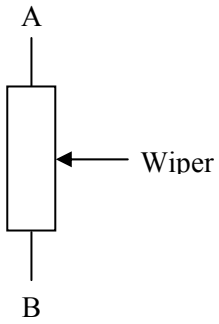
Expressing Resistor Values

Because resistors are very common components, a couple of shortcuts may be taken when writing resistor values. The first is that the "ohm" or Ω abbreviation for the unit may be omitted, so a 10 kΩ resistor may be referred to just as "10k". The second is that the k or M (for kilo and mega respectively) may be written where the decimal point would normally be,

and the decimal point omitted altogether. So a 3,3 kΩ resistor might be written as "3k3", and a 1,5 MΩ resistor as "1M5". The character "R" is also sometimes used in place of the decimal point when there is no scale factor. For example a 1,5 Ω resistor might be written as "1R5".

The Potentiometer

A related component is the potentiometer, which has a variable resistance. This is typically constructed as a circular carbon track with a known resistance and a wiper that can be moved over the track by turning a control knob. The resistance from one side of the track to the other remains constant, but the resistance between either side and the wiper depends on the position of the control knob. The symbol for a potentiometer is shown below.



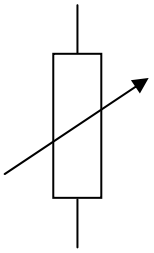
The two ends of the carbon track are represented by the connections at the top and bottom of the diagram. The resistance between these points is fixed. The arrowhead represents the wiper. The three terminals "A", "B" and "W" (for "wiper") are labeled so that they can be referred to in the explanation below. They are not usually labeled in this way.

Let us assume that the potentiometer has a value of 10 kΩ (10 000 Ω). That means that the resistance between A and B is always 10 kΩ. When the wiper is in a central position, as represented in the diagram, then the resistance between A and W would be about half of this – say 5 kΩ, and the resistance between B and W would be the other half of the resistance, also 5 kΩ.

Suppose we turn the control knob so the wiper is closer to A than to B. Then the resistance between A and W would be less than half, say 2 kΩ. The resistance between B and W would be the remainder of the 10 kΩ total resistance, in this case 8 kΩ. Similarly, if we set the wiper all of the way over to B then the resistance between B and W would be 0 Ω (nothing), while the resistance between A and W would be the entire 10 kΩ.

So the resistance between A and W and the resistance between B and W when added together always equal the resistance from A to B, which is the value of the potentiometer.

Potentiometers are often used as controls on electronic equipment, for example the volume control on an audio amplifier or radio receiver. There is also another symbol for a potentiometer:



In this symbol, only the top and bottom lines represent connection points. The line with the arrow point does not represent a separate connection, but rather means that the resistance is variable. This typically represents exactly the same component as the more usual three-terminal symbol shown above. However only two of the terminals are used: one side of the carbon track and the wiper. The other side of the carbon track is left unconnected.

Although the symbols for the potentiometer are drawn vertically, while the symbol for the resistor is drawn horizontally, this was purely for convenience. Any of the symbols, like most electronics symbols, can be drawn either horizontally or vertically.

Summary

The resistor is an electronic component with a defined resistance, tolerance and power rating. The tolerance is the percentage by which the actual resistance may deviate from the nominal value of the resistor. The value and tolerance of resistors is represented using the resistor colour code. The potentiometer is a variable resistor.

Revision Questions

1. A potentiometer is a:
a. Meter.
b. Variable resistor.
c. Battery.
d. Capacitor.

2. How can you determine a carbon resistor's electrical tolerance rating?
a. By using a wavemeter.
b. By using the resistor's colour code.
c. By using Thevenin's theorem for resistors.
d. By using the Baudot code.

3. Which of the resistors below (each identified by its colour coding) would be nearest in value to a 4k7 resistor?
a. Orange violet orange.
b. Yellow green red.
c. Orange violet red.
d. Yellow green orange.

4. What would the colour code be for an 820 Ω resistor, excluding the tolerance band?
a. grey red black.
b. grey red brown.
c. red grey black.
d. red grey brown.

5. What would the value of a resistor with the colour code orange orange orange be?
a. 333 Ω.
b. 3,3 kΩ.
c. 33 kΩ.
d. 330 kΩ.

6. A 10 kΩ resistor has a gold tolerance band. The actual resistance may be:
a. From 9 000 to 11 000 Ω.
b. From 9 500 to 10 500 Ω.
c. From 9 800 to 10 200 Ω.
d. From 9 900 to 10 100 Ω.

7. A 2,2 Ω resistor might be labeled on a circuit diagram as
a. 2k2.
b. 2M2.
c. 2R2.
d. 22R.

8. The label "4M7" on a circuit diagram could refer to:
a. A resistance of 4,7 mega ohms.
b. A current of 4,7 mega amps.
c. A voltage of 4,7 mega volts.
d. Any of the above.

9. The circuit diagram for a resistor is:
a. A straight line.
b. A circle containing a zig-zag line.
c. A rectangle.
d. A triangle.

10. Which of the following types of resistor would not be suitable for radio-frequency applications?
a. A carbon film resistor.
b. A metal film resistor.
c. A wire-wound resistor.
d. A resistor network.

Chapter 5 - Direct Current Circuits

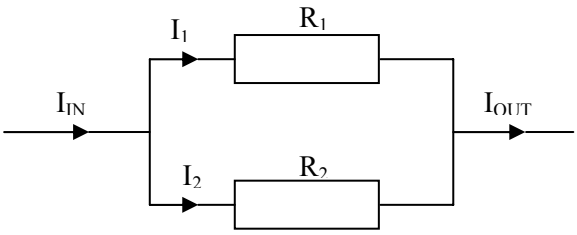
Direct current (abbreviated "D.C.") means a current that is flowing constantly in one direction. It is contrasted to alternating current ("A.C.") like the mains supply, where the direction in which the current flows changes periodically, usually many times every second. Despite the apparent contradiction in terms, it is common practice to speak of a "D.C. voltage" to mean a constant voltage, and an "A.C. voltage" to mean a voltage that is reversing polarity (i.e. exchanging positive and negative terminals) periodically. Although for the moment we shall only consider direct current (D.C.) circuits, when we come to alternating current (A.C.) circuits we will see that the same principles apply

Remember that "voltage" is a commonly used term meaning electric potential, and this will be used in preference to the term "electric potential" for the remainder of these notes, since this is how it is most commonly referred to.

Kirchoff's Laws

Gustav Kirchoff (1824-1887) formalized two very simple laws that allow us to analyze electric circuits. The first is known as Kirchoff's current law.

Kirchoff's Current Law: *At any point in a circuit where two or more wires are joined, the sum of the currents flowing into the point is equal to the sum of the currents flowing away from the point.*



For example, consider the diagram above, which shows two resistors connected "in parallel". The arrows on the lines represent currents. A current I_{IN} flows into the circuit from the left, divides into two currents I_1 and I_2, which flow through resistors R_1 and R_2 respectively. After flowing through the resistors, the currents join again together to give I_{OUT}.

Note that this is not a complete circuit, as we have not shown the source of electric potential that is causing the current to flow. We must assume that there is some voltage source connected so that its positive terminal is connected to the wire on the left hand side of the diagram and its negative terminal is connected to the wire on the right hand side of the diagram in order to make the current flow.

Applying Kirchoff's current law to the point where I_{IN} splits into I_1 and I_2, we see that the sum of the currents flowing into the point – in this case there is only one current, I_{IN} – must equal the sum of the currents flowing out of the point – in this case, $I_1 + I_2$. One way to look at this is that current is a flow of charge, and charge cannot accumulate at a point, so charge must flow out of the point just as fast as it flows in.

In our analogy with a water pipe, if you put a "T" connector on a pipe then the rate at which the water flows out of the two output pipes combined must equal the rate at which the water is flowing into the input pipe, since the water that is coming in has to go somewhere and it cannot accumulate in the T connector.

So in the diagram above we have

$$I_{IN} = I_1 + I_2$$

Referring now to the point where I_1 and I_2 join together to form I_{OUT}, we can again apply Kirchoff's current law which says that the sum of the currents flowing into the point – that is, $I_1 + I_2$ – must equal the sum of the currents flowing out of the point, in this case just I_{OUT}. So this application of Kirchoff's Current Law gives us

$$I_1 + I_2 = I_{OUT}$$

Because both equations have "$I_1 + I_2$" on one side of the equals sign, we can combine them to get

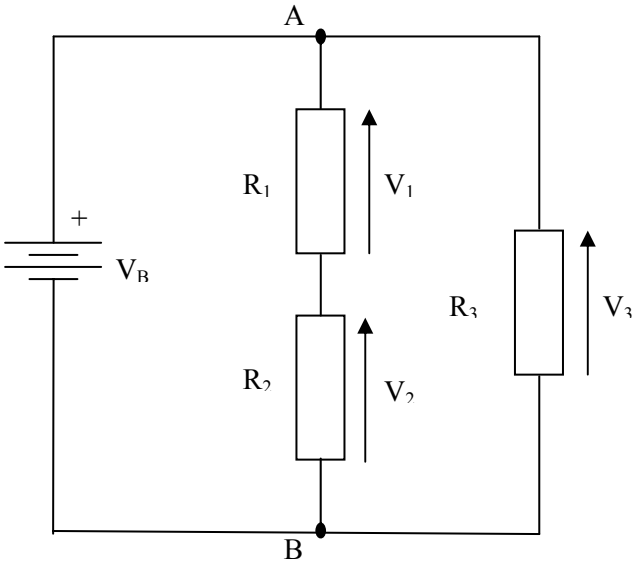
$$I_{IN} = I_{OUT}$$

which makes sense because the charge that is flowing in on the left hand side has to go somewhere, and the only place for it to go is out the right hand side of the diagram.

The second of Kirchoff's laws is Kirchoff's Voltage Law. It can be formulated in two different but equivalent ways. The first formulation, which is the most useful, is as follows.

Kirchoff's Voltage Law (1): *The voltage between any two points in a circuit is equal to the sum of the voltage drops along any path connecting those points.*

This requires some explanation. Consider the circuit below:



The symbol on the left hand side of the diagram represents a battery. The long line always represents the positive terminal, but has been labelled with a "+" sign to make it clear. The battery voltage has also been labelled as V_B. The battery is generating a voltage across R_1 and R_2, which are connected "in series", and across R_3, which is connected "in parallel" with R_1 and R_2.

The voltage applied by the battery will cause a current to flow through R_1 and R_2 and another (possibly different) current to flow through R_3. However we know from Ohm's law that when a current flows through a resistance there will be a voltage across the resistance. The voltage across a resistance is often referred to as a "voltage drop". The voltage drops across R_1, R_2

and R_3 have been labelled as V_1, V_2 and V_3 respectively. The lines with arrowheads are used to indicate what points the voltage drop is across. Note that by convention the arrowhead points towards the positive side, which means that the arrows point in the opposite direction from the direction in which current is flowing in the circuit. (In this circuit, the currents in the resistors are all flowing from top to bottom.)

Voltage Drop: *the potential difference across a component like a resistor caused by the current flowing through the component.*

So what does Kirchoff's Voltage Law tell us about the circuit? Consider points A and B in the diagram. Kirchoff's voltage law tells us that the voltage between points A and B is equal to the sum of the voltage drops along any path connecting A and B. If we call the voltage between A and B "V_{AB}", then applying Kirchoff's Voltage law to the three different paths between A and B gives us:

$V_{AB} = V_B$ (from the path through the battery)
$V_{AB} = V_1 + V_2$ (from the path through R_1 and R_2)
$V_{AB} = V_3$ (from the path through R_3)

In other words, the *same* voltage is found across the battery, across the series combination of R_1 and R_2 and across R_3. Thinking of it in another way, the battery voltage V_B has been applied across both the series combination of R_1 and R_2 and across R_3. The concept is very simple and straightforward, and you should be able to apply it intuitively and hardly ever have to think about its formal statement as Kirchoff's Voltage Law.

At the beginning of the section it was mentioned that there are two different although equivalent formulations of Kirchoff's voltage law. The second is:

Kirchoff's Voltage Law (2): *The sum of the voltage drops around any closed circuit is zero.*

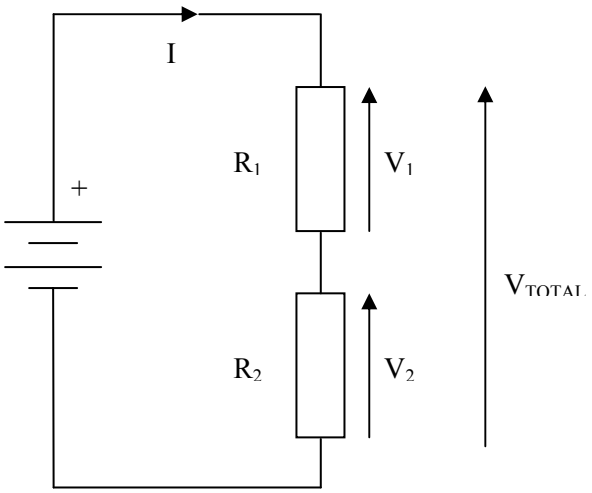
This is somewhat less intuitive than the original formulation. Suppose we take a clockwise trip around the outside circuit in the diagram above, starting and ending at point A. We first go "through" the resistor R_3, and so V_3 is our first voltage drop. Staying on the outside circuit (and so ignoring R_1 and R_2), we next come to the battery. However the voltage across the battery, V_B, is not actually a voltage *drop* because we are moving from the negative terminal to the positive terminal so the voltage is *increasing*. However we can't just ignore it, so we instead count the battery voltage V_B as a *negative* voltage drop and add $-V_B$ to our "sum of voltage drops". Since adding the negative of a number is the same as subtracting that number we get:

sum of voltage drops = $V_3 - V_B$

However we have already seen that V_3 and V_B are equal, so the sum equals zero and Kirchoff is happy!

Resistors in Series

Having mastered Ohm and Kirchoff's laws, we can use these to derive some simple and well-known results. The first is the formula for calculating the effective resistance of two resistors in series. Consider the following circuit:



It shows two resistors connected "in series" so that the same current flows through both of the resistors, although the voltages across each resistor may be different. The current flowing in the circuit is I, while the voltages across R_1 and R_2 are V_1 and V_2 respectively. The voltage across both resistors combined as V_{TOTAL}. The battery is only shown for completeness, to show how the current is being made to flow in the circuit.

Suppose we want to replace the two separate resistors R_1 and R_2 by a single resistor, which will have the same effect. What value of resistor should we choose?

Note that the derivation below is provided for interest only and will not be examined. You only need to know the result that appears in italics at the bottom of this section.

From Ohm's law,

$$V_1 = I R_1$$
and $$V_2 = I R_2$$

From Kirchoff's Voltage Law

$$V_{TOTAL} = V_1 + V_2$$

Replacing V_1 and V_2 in this formula with the values from Ohm's law,

$$V_{TOTAL} = I R_1 + I R_2$$
$$= I (R_1 + R_2)$$

But this is just Ohm's law for a resistor with the value $R_1 + R_2$. In other words, the resistors R_1 and R_2 together behave just as though they were a single resistor with the value $R_1 + R_2$. This gives us the result we are looking for:

When two or more resistors are connected in series, the combined resistance is the sum of the individual resistances.

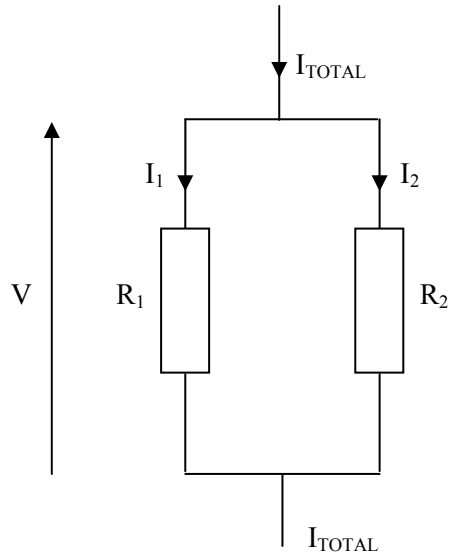
Although we have showed this for two resistors, it is easy to generalize the result to any number of resistors. This is left as an exercise for the reader. (Hint: you don't need Kirchoff's and Ohm's laws, you can just use the result for two resistors and the properties of addition.)

For example, if three resistors with the values 1kΩ, 2kΩ and 4 kΩ were connected in series the combined resistance would be 7kΩ.

Resistors in Parallel

Another way of connecting components is to connect them in *parallel*, so the same voltage appears across each of the components although the currents through them may (and probably will) differ.

Consider the following circuit, which shows two resistors connected in parallel. (This time the source of the potential difference has been omitted – perhaps we should describe it as a "partial circuit".)



The same voltage, V appears across both resistors. The currents through them are I_1 and I_2, while the total current through both resistors combined is I_{TOTAL}.

Once again the derivation is provided for interest only and is not required for the examination.

Using Ohm's law,

$$I_1 = V / R_1$$
and $$I_2 = V / R_2$$

According the Kirchoff's Current Law,

$$I_{TOTAL} = I_1 + I_2$$

Substituting the values of I_1 and I_2 obtained using Ohm's law,

$$I_{TOTAL} = V / R_1 + V / R_2$$

Applying Ohm's law to the whole circuit,

$$V / R_{PARALLEL} = I_{TOTAL}$$
$$= V / R_1 + V / R_2$$

Where $R_{PARALLEL}$ is the equivalent resistance of the two resistors in parallel. Dividing by V,

$$1 / R_{PARALLEL} = 1 / R_1 + 1 / R_2$$

This is the result we were looking for, as it shows the relationship between the value of the combined parallel resistance and the individual resistances. It is not as easy to put into words as it was for resistors in series, but I'll give it a go:

When two or more resistors are connected in parallel, the reciprocal of the equivalent parallel resistance is the sum of the reciprocals of the individual resistances.

(Note: the *reciprocal* of a number is *one divided by* that number.)

Of course, this leaves us with the *reciprocal* of the value we are looking for. Fortunately it is simple to convert the reciprocal of a number back into the number itself – just calculate the reciprocal of the reciprocal and this will be the original number! For example, suppose a 220 Ω resistor is connected in parallel with a 330 Ω resistor. We can find the equivalent combined resistance of the two resistors in parallel as follows:

$$1 / R_{PARALLEL} = 1 / R_1 + 1 / R_2$$
$$= 1/220 + 1/330$$
$$= 0{,}004\ 55 + 0{,}003\ 03$$
$$= 0{,}007\ 58$$

So $R_{PARALLEL} = 1 / 0{,}007\ 58$ (the reciprocal of the reciprocal!)
$= 132\ \Omega$

There is a short cut that can be applied when all the resistances in parallel have the same value. In this special case, if the resistors all have the value R and there are N resistors connected in parallel, then the equivalent resistance is R/N. We leave the proof of this as an exercise for the interested reader.

Practical Example

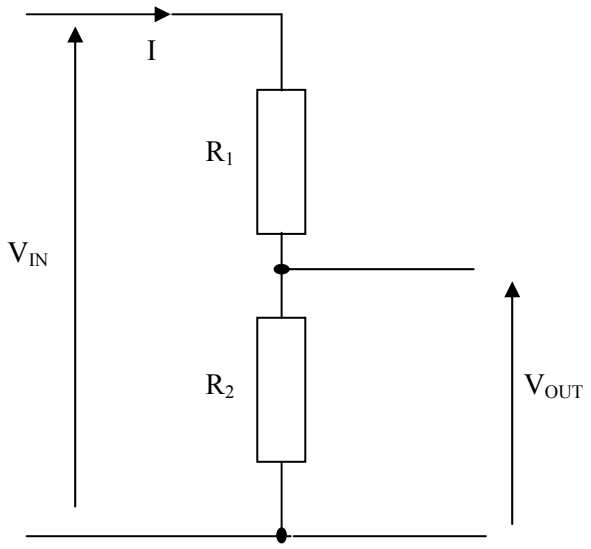
A "dummy load" is a high-powered resistor that can be connected to the antenna port of a transmitter. It enables the transmitter to be tested or aligned without actually having to transmit a signal. Transmitting a signal during testing when not absolutely necessary would cause interference and would be considered extremely bad manners by amateurs.

Commercial dummy loads are available but they are quite expensive. An alternative for the amateur is to make your own. Unfortunately the most commonly available suitable resistors only have a power rating of 2 W, while most transceivers will put out 100 W and would incinerate a 2 W resistor. One solution is to use fifty 2 W resistors all connected in parallel, so that each handles one fiftieth of the transceiver's power. If the resistors are each 2 500 Ω (2k5) then the effective resistance of 50 resistors in parallel is 2500 / 50 = 50 Ω, which is the correct value to match most amateur transceivers.

Remember that you will also want to shield the dummy load to prevent it from inadvertently becoming a fully functional transmitting antenna. This can be achieved by enclosing it in a metal baking powder tin which is chosen because it has a screw-on lid. Drill a hole in the bottom of the tin to accommodate a SO235 (UHF) connector and attach the centre conductor to a piece of stiff wire running down the centre of the tin. Then solder the resistors between this central conductor and the body of the tin. In this way the tin also serves as a heat sink for the resistors as well as a shield for the dummy load.

The Voltage Divider

Two resistors in series can be used as a *voltage divider*. Consider the circuit below:



This shows two resistors connected in series as before. However this time, we are measuring the voltage V_{OUT} across one of the resistors. Our task is to find this output voltage in terms of the input voltage applied across both resistors.

Using our formula for resistors in series, we know that the total combined resistance of R_1 and R_2 in series is $R_1 + R_2$. We can apply Ohm's law to the input voltage and the combined resistance of R_1 and R_2 in series to find the input current I:

$$I = V_{IN} / (R_1 + R_2)$$

Now, if we assume that negligible current is drawn from the output, then the same current I flows through both resistors. Hence we can find the voltage across R_2, which is the output voltage, using Ohm's law:

$$V_{OUT} = I R_2$$

Substituting the value we obtained for I by applying Ohm's law to the series combination of R_1 and R_2 we get

$$\begin{aligned} V_{OUT} &= (V_{IN} / (R_1 + R_2)) R_2 \\ &= V_{IN} R_2 / (R_1 + R_2) \end{aligned}$$

The circuit is known as a "voltage divider" because the output voltage is proportional to but smaller than the input voltage, so the effect of the circuit is to divide the input voltage by a constant (greater than 1).

Summary

Kirchoff's Current Law states that any point in a circuit where two or more wires are joined, the sum of the currents flowing into the point is equal to the sum of the currents flowing away from the point. His Voltage Law states that the voltage between any two points in a circuit is equal to the sum of the voltage drops along any path connecting those points.

We can use these laws in conjunction with Ohm's law to calculate the equivalent values of resistors in series and in parallel. When two or more resistors are connected in series, the combined resistance is the sum of the individual resistances. When two or more resistors are connected in parallel, the reciprocal of the equivalent parallel resistance is the sum of the reciprocals of the individual resistances.

The voltage divider consists of two resistors in series with an output voltage measured across one of the resistors. The formula for the output voltage of a voltage divider is:

$$V_{OUT} = V_{IN} R_2 / (R_1 + R_2)$$

Revision Questions

1 Two 10 kΩ resistors are connected in parallel. If the voltage from a battery across the resistors sets up a current of 5 mA in the one resistor, how much current flows in the second?

a. 10 mA.
b. 2 mA.
c. 20 mA.
d. 5 mA.

2 Two resistors are connected in series to a 9 V battery. The voltage across one of the resistors is 5 V. What is the voltage across the other resistor?

a. 4 V.
b. 5 V.
c. 9 V.
d. 13 V.

3 In a parallel circuit with a voltage source and several branch resistors, what relationship does the total current have to the current in the branch currents?

a. The total equals the average of the branch current in each resistor.
b. The total equals the sum of the branch currents in each resistor.
c. The total decreases as more parallel resistors are added to the circuit.
d. The total is calculated by adding the voltage drops across each resistor and multiplying the sum by the total number of all circuit resistors.

4 Two resistors are connected in series. The combined resistance is 1 200 Ω. If one of the resistors is 800 Ω, what is the value of the other?

a. 1 000 Ω.
b. 800 Ω.
c. 400 Ω.
d. 1 200 Ω.

5 A 100 Ω resistor is connected in series with a 200 Ω resistor. The equivalent resistance of the two resistors is:

a. 100 Ω.
b. 200 Ω.
c. 300 Ω.
d. 400 Ω.

6 A 100 Ω resistor is connected in parallel with a 200 Ω resistor. The equivalent resistance of the two resistors is:

a. 50 Ω.
b. 67 Ω.
c. 75 Ω.
d. 300 Ω.

7 Two light bulbs are connected in series. Which of the following statements is necessarily true:

a. The current flowing through each of the bulbs is identical.
b. The voltage across each of the bulbs is identical.
c. The resistance of each of the bulbs is identical.
d. The light given off by each of the bulbs is identical.

8 Two light bulbs are connected in parallel to the mains. One of them blows, and becomes an open circuit (i.e. no current can flow through it). What will happen to the current flowing through the bulb that is still working.

a. Twice the current as before will flow through the working bulb.
b. No current will flow through the working bulb.
c. The same current as before will flow through the working bulb.
d. Half the current as before will flow through the working bulb.

9 The output voltage from a voltage divider with two equal resistances will be:

a. The same as the input voltage.
b. One quarter of the input voltage.
c. Half the input voltage.
d. One third of the input voltage.

10 A dummy load is made by connecting forty-four 2k2 resistors in parallel. The resistance of the dummy load is:

a. 20 Ω.
b. 50 Ω.
c. 75 Ω.
d. 100 Ω.

Chapter 6 - Power in D.C. Circuits

Power Dissipation in Resistances

When a current flows through a resistance, the resistance will dissipate ("use up") power and generate heat. This principle is used in many electrical devices, for instance in electric bar heaters and kettles, where the elements are just resistances with suitable power handling and heat transfer abilities.

To calculate the power dissipated by a resistance, you multiply the voltage (electric potential) across the resistance by the current flowing through the resistance, so

$$P = V I$$

It is easy to see why. Remember that the electric potential between two points is the amount of energy that it would take to move one unit of charge from the point of lower potential to the point of higher potential. Now that we are allowing the charge to flow from the point of higher potential back to the point of lower potential, this energy is recovered, usually in the form of heat. Since current is the rate of flow of charge, the greater the current the greater the energy that will be given off each second, and hence the greater the power.

For example, suppose an electric kettle draws 5 A at 240 V. Its power consumption is calculated as follows:

$$\begin{aligned} P &= V I \\ &= 240 * 5 \\ &= 1\,200\ W \\ &= 1{,}2\ kW \end{aligned}$$

(Of course kettles usually work off A.C. not D.C. power, but when we get to the section on A.C. power you will see that the same formula applies.)

Using Ohm's Law With The Formula For Power

Of course Ohm's law also deals with voltages and currents (as well as resistances), so it can often be used together with the formula for power. For example, suppose that in the example above we had instead been told that the kettle runs off 240 V and its element has a resistance of 48Ω. We could then use Ohm's law to calculate the current, since

$$\begin{aligned} I &= V / R \\ &= 240 / 48 \\ &= 5\ A \end{aligned}$$

The rest of the calculation would then proceed as above, giving us the same answer of 1,1 kW. Another way is to combine Ohm's law and the formula for power dissipation first, and only bring the actual numbers in at the end.

The formula for power is

$$P = V I$$

But according to Ohm's law, we also know that

$$I = V / R$$

So we can replace the symbol "*I*" in the power equation with "*V / R*" to give

$$P = V\ V / R$$

And since *V* * *V* is just V^2 (pronounced "V squared"), we end up with

$$P = V^2 / R$$

Applying this to the example, where V = 240 V and R is 48Ω, we get

$$\begin{aligned} P &= 240^2 / 48 \\ &= 57\ 600 / 48 \\ &= 1\ 200\ W \\ &= 1{,}2\ kW \end{aligned}$$

Which fortunately is the same answer as before.

In the same way, if you know the current flowing through a resistance and the value of the resistance, but not the voltage across it, then you can use Ohm's law to calculate the voltage across the resistance and then apply the formula for power to calculate the power dissipation. Or these two steps can be combined in a single equation:

$P = V I$ (the formula for power)
and $V = I R$ (Ohm's law)
so $P = I\ I\ R$
$= I^2 R$

This gives you a simple formula for calculating power from current and resistance:

$$P = I^2 R$$

For example, suppose a 50 Ω resistor has a current of 2 A flowing through it. The power dissipated is:

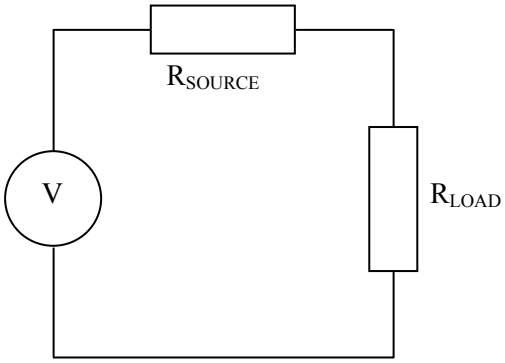
$$\begin{aligned} P &= I^2 R \\ &= 2^2 * 50 \\ &= 4 * 50 \\ &= 200\ W \end{aligned}$$

Exercise

Use Ohm's law to find the voltage across the resistor, and then the formula $P = V\ I$ to calculate the power dissipated by the resistor, and see if you get the same answer.

Matching Source and Load

All real life voltage sources have some *internal resistance*. This can be represented as follows, where the circle with a "V" in it represents a perfect voltage source, R_{SOURCE} is the resistance of the source, and R_{LOAD} is the load resistance. (A *load* is something which the circuit is delivering power to. Depending on the application it might be an antenna, an electric motor, a light bulb or anything else that uses power.)



An interesting question is what load resistance (i.e. what value of R_{LOAD}) will result in the maximum power transfer to the load?

If the load resistance is very low, then a lot of current will flow in the circuit, but the voltage across the load will be small. If the resistance is high, then the voltage across the load will be high, but the current through it will be low. Since $P = V I$ both the current through the load and the voltage across it are important for power transfer.

Although the mathematics is a bit beyond the level of this course, it turns out that the load dissipates the maximum power when the load resistance is exactly equal to the source resistance. In this case, the power dissipated by the load is $V^2 / (4 R_{LOAD})$. This is useful to know when designing power sources such as power amplifiers. You should note, however, that with a matched load the source dissipates just as much power as the load, so heat sinking may be quite important!

Summary

The power dissipated in a resistive load can be found using the formula $P = V I$. This can be combined with Ohm's law to give $P = I^2 R$ and $P = V^2 / R$. In a simple resistor, this power will be dissipated as heat.

All voltage sources have some internal resistance. The maximum power transfer from the source to the load occurs when the load resistance is exactly equal to the source resistance.

Revision Questions

1 A light bulb is rated at 12 V and 3 W. The current drawn when used on a 12 V source is:

a. 250 mA.
b. 750 mA.
c. 4 A.
d. 36 A.

2 The DC current drawn by the final stage of a linear amplifier is 100 mA at 100 V. How much power is consumed?

a. 100 W.
b. 1 kW.
c. 10 W.
d. 1 W.

3 If a power supply delivers 200 W of electrical power at 400 V DC to a load, how much current does the load draw?

a. 0,5 A.
b. 2,0 A.
c. 5 A.
d. 80 000 A.

4 The product of the current and what force gives you the electrical power in a circuit?

a. Magnetomotive force.
b. Centripetal force.
c. Electrochemical force.
d. Electromotive force.

5 A resistor is rated at 10 W. Which of the following combinations of potential difference and current exceeds the rating of the resistor?

a. 2 V, 100 mA.
b. 20 V, 200 μA.
c. 1 kV, 25 mA.
d. 10 mV, 2 A.

6 The starter motor of a motor car draws 20 A from the 12 V battery. How much power does it use?

a. 2,4 W.
a. 24 W.
b. 240 W.
c. 2,4 kW.

7 What is the resistance of the motor in question 6?

a. 0,6 Ω.
b 1 Ω.
c 6 Ω.
d 10 Ω.

8 The internal resistance of a car battery is found to be 0,2 Ω. Into what load resistance will it deliver the maximum power?

a. 0,1 Ω.
b. 0,2 Ω.
c. 0,6 Ω.
d. 1,2 Ω.

9 At its peak, a lightning bolt has a voltage of 100 million volts and 10 000 A. How much power does it deliver?

a. 10^9 W.
b. 10^{10} W.
c. 10^{11} W.
d. 10^{12} W.

10 A current of 2 mA is measure in a 1 kΩ resistor. How much power is the resistor dissipating?

a. 2 mW.
b. 4 mW.
c. 2 W.
d. 4 W.

Chapter 7 - Alternating Current

Introduction

In direct current (D.C.) circuits, the current always flows in one direction. This is because the two terminals of the voltage sources used to power these circuits always have the same polarity – one terminal (the positive one) is always positive with respect to the other terminal. This causes the current to flow in only one direction in the circuit.

However in other circuits, the direction in which the current flows is constantly changing. The current flows first in one direction, then in the reverse direction, then in the original direction again and so on, with the direction changing at regular intervals, usually many times each second. The circuits are called *alternating current* (A.C.) circuits. Power for these circuits may be supplied by alternating current (A.C.) power supplies, such as the mains supply. With A.C. power supplies, there is no "positive" or "negative" terminal. Instead, one terminal will be positive with respect to the other for a brief period, and then the roles will reverse and the other terminal will become more positive for a brief period, and so on. Although the abbreviation A.C. stands for "alternating current", it is also used to refer to voltages, in phrases such as "An A.C. Voltage" and "15 VAC.".

The Sine Wave

If you were to plot the voltage or current in an A.C. circuit against time, there are many possible shapes (known as "waveforms") that this could take. For example:

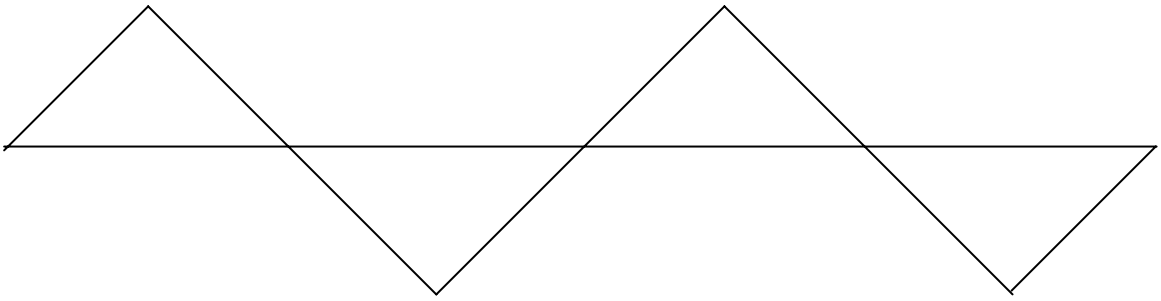


Figure 1: A Triangular Wave

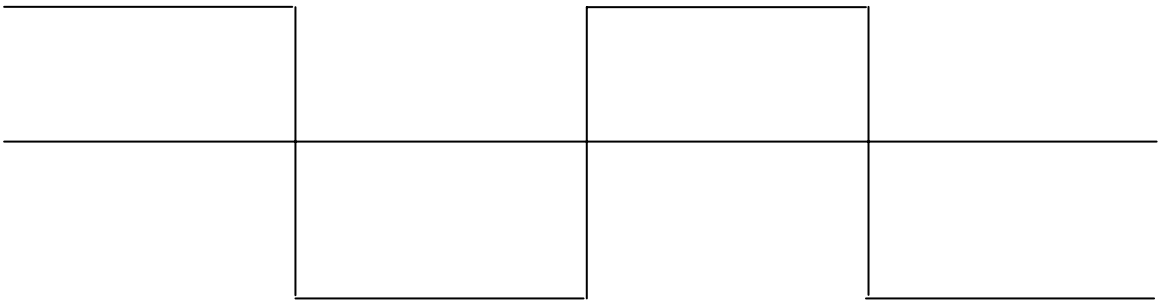


Figure 2: A Square Wave

However, when we analyze A.C. circuits, we normally think of the waveform as being a "sine wave". This is a waveform given by the mathematical equation:

$$V = V_{PEAK} \sin(2\pi ft)$$

Where V_{PEAK} is the peak voltage of the waveform, f is its frequency, t is time, π is the mathematical constant "pi" (approximately 3,14) and *sin* is the trigonometric "sine" function. The shape of a sine wave is shown below. Note that it is *not* two semi-circles, which is how it is sometimes incorrectly drawn.

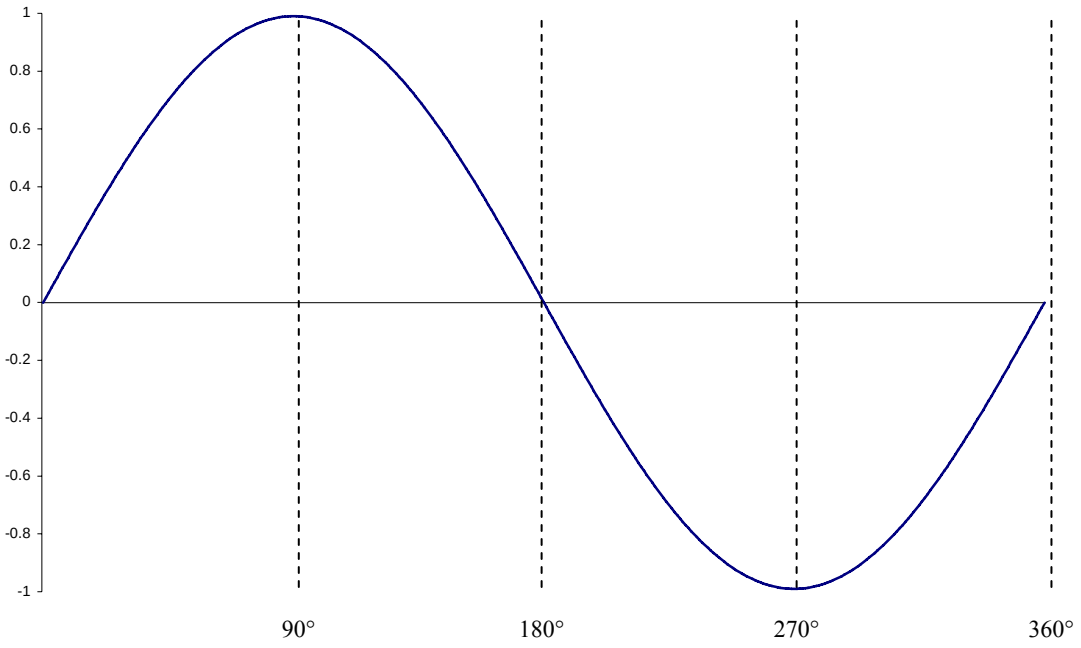


Figure 3: A Sine Wave

The reason why we deal mostly with sine waves in circuit analysis is because the French mathematician Joseph Fourier (1768-1830) showed that any other waveform could be decomposed into a number of sine waves of different frequencies. So if we know how a circuit responds to a sine wave then we can easily calculate its response to any other waveform using the technique known as *Fourier analysis*. A sine wave represents a "pure" A.C. waveform that contains only a single frequency, known as the *fundamental*. Any other waveform includes both the fundamental and *harmonics*, which are integral multiples of the fundamental frequency.

Cycles and Half Cycles

An A.C. waveform consists of many identical *cycles* one after another. Figure 3 shows one complete cycle of a sine wave, while Figure 1 shows two complete cycles of a triangular waveform.

Question: *How many cycles of a square wave are shown in Figure 2?*

Usually electrical waveforms like A.C. voltages and currents are positive for half the time and negative for the other half. When we want to refer just to the positive or negative period, we speak of the "positive half cycle" and "negative half cycle".

Period and Frequency

The period of a waveform is the time taken for one cycle, which is usually expressed in seconds, milliseconds or microseconds.

Definition: *The period of a waveform is the time taken for one complete cycle.*

The frequency of a waveform is the number of cycles per second. The unit of frequency, one cycle per second, is called the Hertz (abbreviated Hz) in honour of the German physicist Gustav Hertz (1887-1975).

Definition: *The frequency of a waveform is the number of cycles per second.*

Since period is the number of seconds per cycle, and frequency is the number of cycles per second, it follows that the period and frequency of a waveform are reciprocals of each other. That is:

$$t = 1/f$$
and $$f = 1/t$$

where t is the period (in seconds) and f the frequency in Hertz.

For example, the mains frequency in South Africa is 50 Hz (50 cycles per second). The period can be found from:

$$t = 1/f$$
$$= 1/50$$
$$= 0{,}02 \text{ s}$$
$$= 20 \text{ ms}$$

Wavelength and the Speed of Light

Electrical currents and voltages move through wires at the speed of light, which is a very high but not infinite speed. Radio waves transmitted from an antenna also travel at the speed of light. The speed of light, which is usually represented by the symbol c in physics, is approximately $3 * 10^8$ m/s. That is just over a billion (10^9) kilometers per hour!

Think about an A.C. waveform with a constant frequency moving through a very long wire at the speed of light. The start of one cycle will occur at a particular point in time, and hence at a particular distance along the wire. The start of the next cycle will occur a certain time later (this time difference being the *period* of the wave), during which the wave, which is traveling at the speed of light, will have moved some distance further along the wire. Since the speed of light is constant, and the time between successive cycles of the wave (the period) is also constant, the distance traveled by the wave between the start of one cycle and the start of the next must also be constant (for this particular wave). This distance is known as the *wavelength* of the wave.

Definition: *The wavelength of a wave is the distance it travels in one cycle.*

Because wavelength represents the distance the wave travels during a certain time, it is related to the period and frequency as follows:

$$\lambda = c\,t$$
and $$\lambda = c / f$$

where λ is the wavelength in meters, c the speed of light in meters per second, t the period in seconds and f the frequency in Hertz.

For example, one of my favourite radio stations is Cape Talk, which broadcasts on a frequency of 567 kHz. The corresponding wavelength can be calculated as follows:

$$\lambda = c / f$$
$$= 3 * 10^8 / (567 * 10^3)$$
$$= 529\ m$$

This is the distance that the radio waves transmitted by Cape Talk will travel during one complete cycle.

There is a short cut that is quite useful for radio amateurs. Because we express most of our frequencies in megahertz (millions of cycles per second), you can avoid having to deal with lots of zeros (or with scientific notation) by using the formula:

$$\lambda = 300 / F$$

where λ is the wavelength in meters and F the frequency in MHz. For example, a frequency of 14,100 MHz has a wavelength of:

$$\lambda = 300 / F$$
$$= 300/14,1$$
$$= 21,3\ m$$

Note that the higher the frequency, the shorter the wavelength and vice-versa. You can also calculate the frequency from the wavelength using the formula:

$$F = 300 / \lambda$$

where F is the frequency in MHz and λ the wavelength in meters. For example, the amateur "two-meter" band has frequencies of approximately:

$$F = 300 / \lambda$$
$$= 300 / 2$$
$$= 150\ MHz$$

The actual frequencies of the two-meter band in South Africa are 144-146 MHz. The reason for the discrepancy is that "two-meter band" is intended as a name for the band, not an accurate representation of its wavelength. "Two meter and seven centimeter band" would be a bit of a mouthful!

Phase

It is possible to have two sine waves of the same frequency but where the cycles start at different times. In this case, we talk of the waves having a *phase difference*. The phase difference is usually expressed in degrees. One complete cycle has 360º so for example a phase difference of one quarter of a cycle would be 90º.

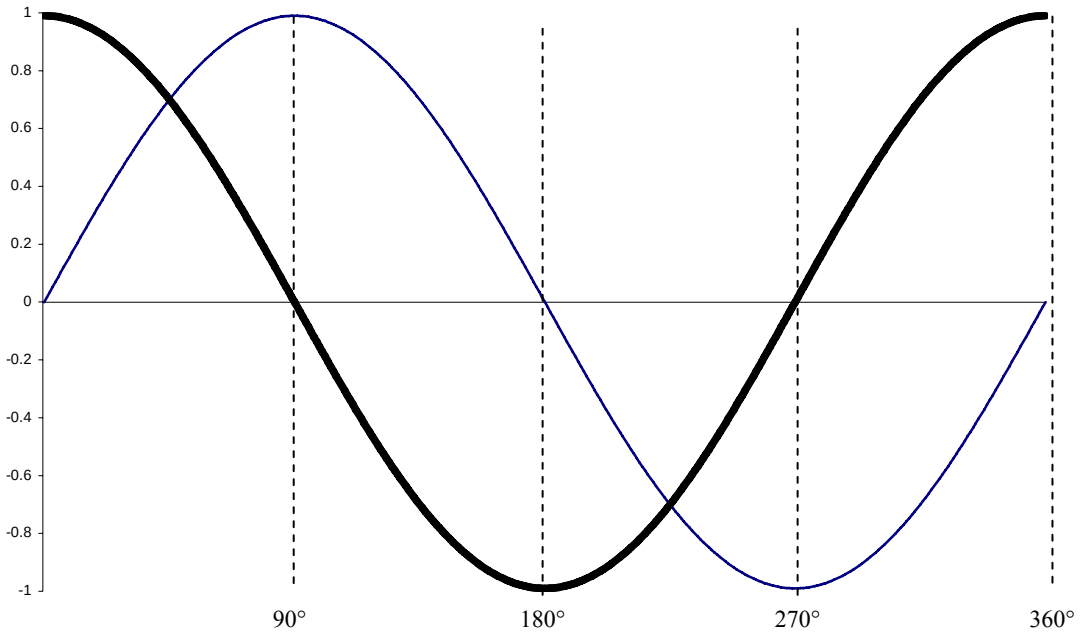


Figure 4: Two sine waves with a phase difference of 90°

The wave that reaches a certain part of its cycle – for example, the maximum positive value – before the other is said to *lead* the other wave. Conversely, the wave that reaches that part of its cycle after the other wave is said to *lag* the other wave. In Figure 4, the wave drawn with a thick line *leads* the other wave by 90º because it gets to its maximum positive value *before* the other wave does.

RMS Voltage and Current

Remember the formulae to find power dissipation given the value of a resistance and the voltage across the resistance:

$$P = V^2 / R$$

If this formula is applied to sine wave, one can see that the power dissipation is at a maximum at the positive and negative peaks of the wave, and at a minimum when the voltage is zero. (Remember that the square of a negative number is a positive number.)

If we were able to average out the square of the voltage through a full cycle of the sine wave, we could calculate an equivalent D.C. voltage that would cause the same power dissipation in a resistor. This is known as the "root mean square" or R.M.S. voltage. ("Mean" is a mathematical term for "average".)

Definition: *The R.M.S. value of an A.C. voltage is the value of the D.C. voltage that would cause the same power dissipation in a resistance.*

For a sine wave, the R.M.S. voltage is the peak voltage (the maximum voltage reached on both positive and negative peaks) divided by the square root of two.

$$V_{RMS} = V_{PEAK} / \sqrt{2}$$

so $$V_{RMS} = 0,707\ V_{PEAK}$$

Note that this formula only works for a sine wave.

Whenever one gives the value of an A.C. voltage, the value is assumed to be the R.M.S. value unless otherwise noted. For example, the mains voltage in South Africa is specified as 240V AC. This is the *RMS* value. We can calculate the peak value as follows:

$$V_{RMS} = V_{PEAK} / \sqrt{2}$$
so $$V_{PEAK} = \sqrt{2}\, V_{RMS}$$
$$= 1{,}41 * 240$$
$$= 338\ V$$

In the same way, A.C. current is usually expressed as an R.M.S. current unless otherwise specified. The definition is similar:

Definition: *The R.M.S. value of an A.C. current is the value of the D.C. current that would cause the same power dissipation in a resistance.*

The R.M.S. current can be found from the peak current using a similar formula to the one used to find the R.M.S. voltage from the peak voltage:

$$I_{RMS} = I_{PEAK} / \sqrt{2}$$
so $$I_{RMS} = 0{,}707\ I_{PEAK}$$

The nice thing about working with R.M.S. voltages and currents is that Ohm's law and the formulae for power work for A.C. voltages and currents just like they do for D.C. voltages and currents, as long as you use the R.M.S. values.

For example, if the element of a kettle that runs of 240 V A.C. (R.M.S.) has a resistance of 48Ω, then the current flowing through the element is

$$I = V / R$$
$$= 240 / 48$$
$$= 5\ A\ (R.M.S.)$$

Although We have noted that the 5 A is an R.M.S. value, this would not normally be necessary as R.M.S. measurements are assumed for all A.C. values unless otherwise specified.

Similarly the power can be calculated using the usual formula,

$$P = V I$$
$$= 240 * 5$$
$$= 1{,}2\ kW$$

Because we are using R.M.S. values, the standard formula gives the right answer.

Summary

A.C. waveforms consist of many identical cycles, one after another. Sine waves consist of a single frequency known as the *fundamental*. All other waveforms have additional frequencies, the *harmonics*. The period of a waveform is the time taken for one complete cycle. The frequency of a waveform is the number of cycles per second. The wavelength of a wave is the distance it travels in one cycle. The wavelength and frequency of a wave are related by the formula:

$$F = 300 / \lambda$$

where F is the frequency in MHz and λ the wavelength in meters. Phase differences are expressed in degrees, with 360º in one complete cycle.

A.C. voltages and currents are expressed as R.M.S. values. The R.M.S. value of an A.C. voltage or current is the value of the D.C. voltage or current that would cause the same power dissipation in a resistance. The R.M.S. voltage can be calculated from the peak voltage using the formula

$$V_{RMS} = ,707\ V_{PEAK}$$

Revision Questions

1 The frequency of an AC waveform is defined in the unit:
a. Seconds.
b. Velocity.
c. Period.
d. Hertz.

2 The frequency of 5 Hz has a period of:
a. 2 seconds.
b. 300 seconds.
c. 0,2 seconds.
d. 1,2 seconds.

3 The wavelength of a signal of 100 MHz in free space is:
a. 30 mm.
b. 0,3 m.
c. 3,0 m.
d. 30,00 m.

4 A wave has a period of 20 ms. Its wavelength in free space is:
a. 6 km.
b. 60 km.
c. 600 km.
d. 6 000 km.

5 Two sine waves are 180º out of phase. When the one wave is at its maximum positive value the other:
a. Is also at its maximum positive value.
b. Is at its most negative value.
c. Is at zero.
d. Cannot be determined from the information given.

6 Which waveform consists of only the fundamental frequency, without any harmonics:
a. A square wave.
b. A sine wave.
c. A triangular wave.
d. A saw-tooth wave.

7 Which value represents the ratio of RMS to Peak value of an AC waveform?
a. 0,5.
b. 0,636.
c. 1,414.
d. 0,707.

8 What is the value of an AC waveform, representing the equivalent heating effect to a DC voltage, known as ?
a. RMS value.
b. Average value.
c. Peak value.
d. Corrected value.

9. The mains voltage in the U.S.A. is 115 V RMS. What is the peak voltage?
a. 81 V.
b. 115 V.
c. 163 V.
d. 220 V.

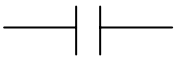
10. The mains voltage in South Africa is 240 V RMS. If this is applied across a heating element with a resistance of 576 Ω, how much power will be dissipated?
a. 10 W.
b. 57,6 W.
c. 100 W.
d. 576 W.

11. An electric geyser operating from the 240 V AC RMS mains supply consumes 2,4 kW. What current does it draw?
a. 10 A RMS.
b. 10 A peak.
c. 10 A average.
d. 10 A DC.

12. A hi-fi loudspeaker has a resistance of 8 Ω. When it is delivering 8 W, what is the RMS voltage across the speaker?
a. 1 V RMS.
b. 8 V RMS.
c. 10 V RMS.
d. 80 V RMS.

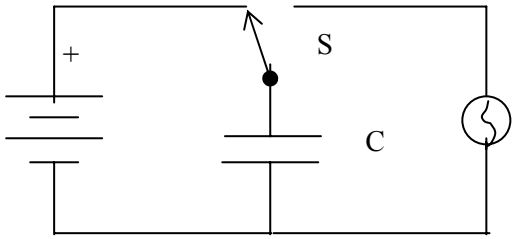
Chapter 8 - Capacitance and the Capacitor

The capacitor is a component that consists of two electrically conductive *plates* separated by a thin layer of some insulating material known as the *dielectric*. The circuit symbol for a capacitor is quite suggestive of its construction:



The two vertical lines represent the conductive plates and the gap between them represents the insulating dielectric.

Capacitors have a property known as *capacitance*, which is the ability to store energy in an electric field between the plates. To see how this works, consider the circuit below:



This shows a capacitor connected to a switch that can either be used to connect it to the battery on the left or to the light bulb on the right.

Let us start by thinking what happens in terms of electrons. When the capacitor is connected to the battery, some of the negatively charged electrons in the upper plate of the capacitor are attracted towards the positive terminal of the battery. At the same time, some electrons flow from the negative terminal of the battery to the lower plate of the capacitor. In effect the battery is acting as an “electron pump”, pumping some electrons from the top plate of the capacitor, through the battery and to the bottom plate of the capacitor.

Through this process the upper plate of the capacitor loses some of its electrons, so it becomes positively charged. At the same time the lower plate gains some excess electrons and so becomes negatively charged. In terms of conventional current, a current flows through the capacitor from top to bottom, which generates a potential difference across the capacitor, with the upper plate becoming more positively charged and the lower plate becoming more negatively charged. This process is known as “charging” the capacitor.

The voltage that is developed across the capacitor opposes the flow of current through the capacitor. As the voltage across the capacitor increases the current through it decreases and when the voltage across the capacitor is equal to the battery voltage the current stops flowing altogether. The capacitor is now fully charged.

Assume that the switch is now flipped so that the capacitor is connected to the light bulb. The excess of electrons on the negatively charged lower plate will be attracted to the positively charged upper plate, which has a shortage of electrons, so a current will flow – in conventional terms, the current flows from the positively charged upper plate to the negatively charged lower plate. This current will make the light bulb glow. As the current flows, the charges on the capacitor plates will gradually return to normal, and the potential

difference across the plates will reduce. This will reduce the current flowing in the circuit until eventually both plates have the same number of electrons so there is not potential difference across the capacitor and the current will stop flowing altogether. This process is known as "discharging" the capacitor.

When a capacitor is charged in a D.C. circuit it has a voltage across it and a current flowing through it, so power is being dissipated in the capacitor according to the formula $P = V I$. However this power is not being converted into heat as it was in a resistor. Instead, energy is being stored in the electric field between the plates. When the capacitor is discharged this energy is released – in this case, it causes the light bulb to glow.

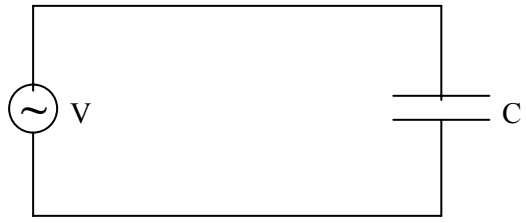
Capacitors come in different values. The value of a capacitor (its capacitance) depends on:

- The surface area of the plates. The greater the area, the greater the capacitance.
- The distance between the plates. The greater the distance, the lower the capacitance.
- A property of the dielectric called its *dielectric constant*.

Large capacitors (meaning those with high capacitance, not necessarily related to their physical size) are able to store a lot of energy by allowing a large excess of positive or negative charge to accumulate on the plates. Small capacitors can only store a little energy, as only a small amount of excess charge can be accumulated. The value of a capacitor is measured in Farads, and typical practical capacitors range in size from 1 pF to 1 000 μF or so.

Capacitors in A.C. Circuits

Capacitors get more interesting in A.C. circuits. Consider this circuit, which shows an A.C. voltage source connected to a capacitor through a resistor:



Note that the "~" symbol on the voltage source means that it is an A.C. source. The symbol "V" represents the voltage of the source, and "C" represents the value of the capacitor.
The first question is whether a current will flow at all. If the voltage source was D.C. then the capacitor would soon charge up to the same voltage as the voltage source, and no more current would flow (except for a very small *leakage* current). However because in this circuit we have an A.C. voltage source, the situation is different. As the current flows in one direction, the capacitor will begin to charge up and the potential difference this causes will oppose the flow of current through the capacitor. However when the current changes direction, the capacitor will start to discharge and the energy it had "borrowed" will be returned to the circuit. Eventually the capacitor will be fully discharged and will start to charge up again but with the reverse polarity. Then when the current direction reverses again, the capacitor can discharge again before once again charging in the original direction.

So with an A.C. source a current will flow through a capacitor. It is interesting to consider the effect of frequency. A low frequency A.C. source will cause the current to flow for a long time in one direction. During this time, the capacitor will become appreciably charged and the potential difference that forms across its plates will significantly oppose the flow of current in the circuit. So for a low frequency source, only a small current will flow. On the other hand, with a high frequency source current will only flow in one direction for a short time before

reversing direction. This will not be long enough to charge the capacitor much, so not much potential difference will develop across the plates, and there will not be much opposition to the flow of current. So for a high frequency source, a larger current will flow.

Reactance

The opposition to the flow of current that we experience with capacitors in an A.C. circuit is not resistance. If it were resistance, then power would be dissipated by the capacitor. However we have seen that the energy that is "borrowed" during one half cycle is returned to the circuit during the next half cycle. The opposition to the flow of current in a capacitor is called "reactance" and usually given the symbol X. The formula for the reactance of a capacitor is:

$$X_C = -1 / (2 \pi f C)$$

Where X_C is the capacitive reactance in ohms, f the frequency in Hertz and C the capacitance in Farads. Note that the reactance *decreases* as the frequency *increases*. This is because capacitors oppose the flow of current less at high frequencies.

Our original question was how much current will flow in the circuit. Fortunately, Ohm's law works for reactances in just the same way as it does for resistances:

$$I = -V / X$$

So once we have calculated the reactance of the capacitor, we can easily calculate the current flowing in the circuit. However note that although resistance and reactance are both measured in ohms, you cannot add them together – they are different quantities.

For example, in the circuit above suppose the voltage V is 1 V, the capacitance of the capacitor is 1 nF (10^{-9} F) and the frequency is 1 MHz (10^6 Hz). Then the reactance of the capacitor is:

$$\begin{aligned} X_C &= -1 / (2 \pi f C) \\ &= -1 / (2 * 3{,}14 * 10^6 * 10^{-9}) \\ &= -1 / (0{,}006\ 28) \\ &= -159\ \Omega \end{aligned}$$

Don't worry too much about the minus sign in the equations. Capacitive reactances are always negative. The current flowing in the circuit can be found using Ohm's law in a slightly modified form:

$$I = V / |X|$$

Here, |X| means "the magnitude of X", in other words the value of X but without the minus sign if it has one. So

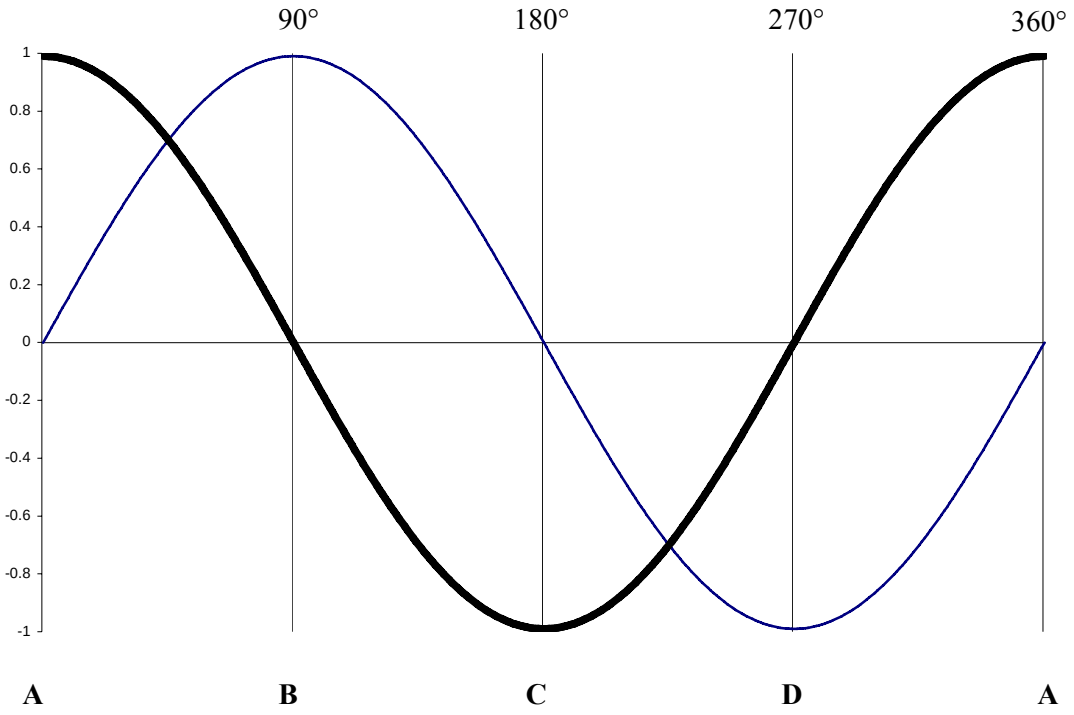
$$\begin{aligned} I &= 1 / 159 \\ &= 0{,}006\ 3\ A \\ &= 6{,}3\ mA \end{aligned}$$

Phase of Current and Voltage

The current flowing through a capacitor and the voltage across the capacitor have an interesting property: they are always 90º out of phase. To be precise, the current flowing through a capacitor *leads* the voltage across the capacitor by 90º, so the voltage across the

capacitor *lags* the current flowing through the capacitor by 90º. In the graph below, the thick line represents the current through the capacitor, while the thin line represents the voltage across the capacitor.

(The material from here until the next major heading is optional and will not be examined.)



This is not as counterintuitive as it might seem at first glance. Remember that the capacitor is charging – that is, the voltage across its plates is increasing – for as long as there is current flowing through it in the right direction. So the capacitor should reach maximum positive charge – i.e. with the maximum positive voltage across the plates – when a positive current has been flowing through it for as long as possible. This is exactly what happens at the point in time labelled "B" above. Similarly, it should reach maximum negative charge at the point where a negative current has been flowing through it for as long as possible, which it does at "D".

Also, since the rate at which a capacitor charges or discharges depends on the current flowing through it, this rate should be greatest at the points of maximum current. For example, at "A" where the maximum positive current is flowing through the capacitor, the rate at which it is charging is greatest. Similarly at "C", where the maximum negative current flows through the capacitor, is where its rate of discharge is greatest.

This is another way in which reactance differs from resistance. The voltage across a resistance is always in phase with the current through the resistance, while the voltage across a reactance is always 90º out of phase to the current flowing through the reactance. In fact, this explains why there is no power dissipated by a reactance. The formula for power is:

$$P = V I$$

However remember that a positive number multiplied by another positive number or a negative number multiplied by a negative number both give a positive result; while a positive number multiplied by a negative number or vice-versa gives a negative result.

If you look at the graph above showing the voltage across and current through a capacitor, you will see that in the first quarter of the graph from "A" to "B", both voltage and current are positive, so the power "dissipated" is positive. However in the last quarter of the graph, between "D" and "A", the voltages and currents have exactly the same values (although in reverse), but this time the voltage is negative while the current remains positive, so the overall result is negative. This precisely cancels out the positive power dissipation in the first quarter of the graph.

Similarly, between "C" and "D" the voltage and current are both negative, so the result is a positive power "dissipation". However the voltage and current have exactly the same values between "B" and "C" (again in reverse), but this time the voltage is positive while the current remains negative, so the overall power "dissipation" is negative and exactly cancels out the positive power dissipation between "C" and "D".

So the positive dissipation from "A" to "B" and from "C" to "D" is exactly cancelled out by the negative "dissipation" from "B" to "C" and from "D" to "A". This is just a reflection that the capacitor is "borrowing" energy as it charges, only to "return" it as it discharges.

(End of optional material.)

Capacitors in Parallel and Series

Two or more capacitors connected in parallel are equivalent to a single capacitance with a value equal to the sum of the values of the individual capacitors.

So for capacitors connected in parallel,

$$C_{EQUIV} = C_1 + C_2 + \ldots$$

Note that this is similar to the equation for resistors in *series*.

For capacitors connected in series,

$$1/C_{EQUIV} = 1/C_1 + 1/C_2 + \ldots$$

Note that this is similar to the equation for resistors in *parallel*.

Types of Capacitor

Like resistors, capacitors come in several different types that are designed for different applications.

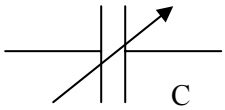
- Ceramic capacitors are generally good for radio frequency (RF) applications and are inexpensive but their tolerance is poor (around ±10%) so they should not be used in critical applications such as the frequency determining elements in oscillators or filters. They are available in values ranging from 100 pF to 100 nF or so, and in high voltage ratings up to 15 kV.

- Silvered Mica capacitors are also good at RF and have much higher tolerances (typically ±1%) but are quite expensive. They are only available in fairly small values from 1 pF to 100 nF.

- Polycarbonate capacitors are suitable when higher capacitance values are required at medium tolerances (±5% is typical). Values range from 10 nF to 10 μF.

- Electrolytic capacitors use metal (usually aluminium) foil as one "plate" of the capacitor and a conductive fluid as the other "plate". The insulating dielectric is a very thin chemical layer that is deposited on the metal film by the dielectric fluid. Electrolytic capacitors can have very high values, up to 100 F, but most of them are *polarised* meaning that one of the terminals must always be positive with respect to the other. This makes them most suited to D.C. applications like power supplies.

- Variable capacitors consist of two sets of plates. Turning the control knob moves one of the sets of plates and varies how much they overlap the other, fixed, plates. In this way the capacitance can be varied. Variable capacitors are often used as the tuning controls of radios. (However modern digital transceivers usually use a digital control called a *shaft encoder* instead.) The symbol for a variable capacitor is shown below:



Summary

Capacitors consist of two conducting plates separated by an insulating dielectric. Capacitors have a property known as *capacitance*, which is the ability to store energy in an electric field between the plates. The energy is stored when the capacitor is *charged* and released when it is *discharged*. The capacitance of a capacitor depends on the surface area of the plates, the distance between the plates and the dielectric constant of the insulating material.

In A.C. circuits capacitors exhibit *reactance* which opposes the flow of current. Although reactance is measured in ohms, it is not the same as resistance as no energy is being dissipated. The reactance of a capacitor is given by the formula:

$$X_C = -1 / (2 \pi f C)$$

Ohm's law can be applied using the magnitude of a reactance in place of a resistance

$$V = I |X| \quad \text{or} \quad |X| = V / I \quad \text{or} \quad I = V / |X|$$

The current flowing through a capacitor *leads* the voltage across the capacitor by 90º. Conversely, the voltage across a capacitor *lags* the current flowing through the capacitor by 90º.

For capacitors connected in parallel,

$$C_{EQUIV} = C_1 + C_2 + \ldots$$

While for capacitors in series,

$$1/C_{EQUIV} = 1/C_1 + 1/C_2 + \ldots$$

There are many different types of capacitor suited to different purposes. Electrolytic capacitors are usually polarised, and one terminal must always remain positive with respect to the other. *Variable capacitors* are used as the tuning control in many radios.

Revision Questions

1 The phase shift between voltage and current in a capacitor is:
a. 90 degrees.
b. 45 degrees.
c. 360 degrees.
d. In phase.

2 Three capacitors of 1 µF are connected in parallel. The equivalent capacitance is:
a. 0,33 µF.
b. 3,0 µF.
c. 0,3 µF.
d. 33,33 µF.

3 A capacitor of 250 pF is required to resonate a tuned circuit. A 100 pF capacitor is connected in parallel to a variable capacitor. What value must the variable capacitor be set to achieve resonance?
a. 150 pF.
b. 300 pF.
c. 350 pF.
d. 400 pF.

4 A value of 1 000 pF is equal to:
a. 10 nF.
b. 1 nF.
c. 0,1 nF.
d. 100 nF.

5 The energy in a charged capacitor is stored in the:
a. Voltage across the terminals.
b. Current applied to the capacitor.
c. The electric field between the plates.
d. Form of magnetism.

6 The unit of capacitance is called?
a. Farad.
b. Permeability.
c. Conductance.
d. Impedance.

7 What is the total capacitance of two or more capacitors connected in parallel?
a. The same as either capacitor.
b. Half the capacitance of either capacitor.
c. Twice the capacitance of either capacitor.
d. The capacitance cannot be determined without knowing the exact values of the capacitors.

8 What do the units microfarad and picofarad specify?
a. Inductance.
b. Capacitance.
c. Resistance.
d. Current.

9 As the plate area of a capacitor increases, its capacitance:
a. Decreases.
b. Increases.
c. Stay the same.
d. Becomes voltage dependent.

10 Which of the factors below would NOT influence the capacitance value of a capacitor?
a. Area of the plates.
b. Distance between the plates.
c. Voltage rating.
d. Di-electric constant of the material between the plates.

11 The (magnitude of the) reactance of a capacitor:
a. Remains constant with changing frequency.
b. Increases with increasing frequency.
c. Decreases with increasing frequency.
d. Increases with decreasing frequency.

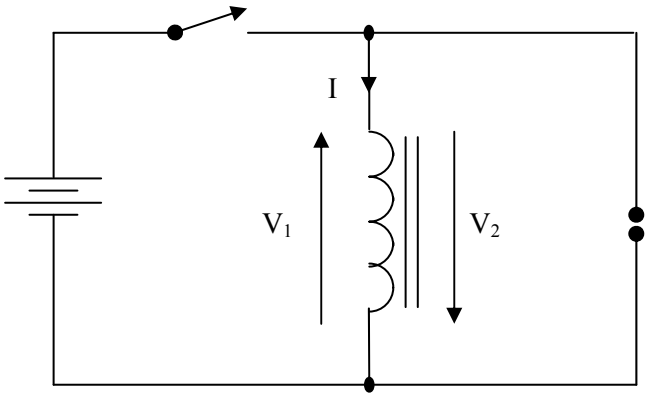
12 The capacitive reactance of a 16 µF, 40 V working electrolytic capacitor to a signal of 100 Hz is:
a. 1 000 Ω.
b. 10 kΩ.
c. 10 Ω.
d. 100 Ω.

13 If the alternating frequency applied to a capacitor is doubled, the capacitor's capacitive reactance will be:
a. doubled.
b. four times original value.
c. one quarter the original value.
d. halved in value.

Chapter 9 - Inductance and the Inductor

A typical inductor consists of a coil of wire, which may be wound around a former or may be self-supporting. When a current flows through the wire, it generates a magnetic field, just like an electromagnet would. Whenever the current flowing through the inductor changes, the corresponding changes to the magnetic field induce a voltage into the inductor that opposes the change in the flow of current. This is known as "self inductance" since the voltage is induced in the same coil that generates the magnetic field.

For example, consider the following circuit:



In this circuit, a battery is connected via a switch to an inductor. (The inductor is the component in the middle of the diagram that looks like a coil of wire). A spark gap is connected in parallel with the inductor (it is represented by the two dots on the right hand side of the diagram).

When the switch is closed, there is initially no current flowing in the inductor. However the potential difference of the battery applied across the battery will cause a current to flow. This causes the inductor to generate a magnetic field, and the growing magnetic field induces a voltage V_1 into the inductor that opposes the attempt to increase the current through the inductor. This means that the current I flowing through the inductor will grow only gradually, rather than reaching its full value as soon as the switch is closed.

When the switch is opened, the magnetic field starts to collapse, which induces a voltage V_2 across the inductor. V_2 acts to oppose the reduction in I that was initiated by opening the switch. Because there is no low-resistance path around the circuit with the switch opened, the only way it can do this is to generate a voltage that is high enough to cause a spark to jump across the spark gap. This induced voltage across an inductor, which is also called the *back EMF*, may be many times the supply voltage.

This is essentially how the ignition circuit in cars with old (non-electronic) ignition systems works. The ignition coil is an inductor, and the points act as a switch that opens, cutting off the current supply to the ignition coil and causing it to generate a high back EMF across one of the spark plugs. In this way a 12 V battery can generate a voltage of several thousand volts across the spark plug.

Another way of looking at this is that when the switch is closed and current flows setting up a magnetic field, energy is taken from the circuit and stored by the inductor in its magnetic field. When the switch is opened, the inductor returns that energy to the circuit as its magnetic

field collapses. So like a capacitor, an inductor "borrows energy from" and "returns energy to" the circuit, but does not actually dissipate power.

Inductor Values

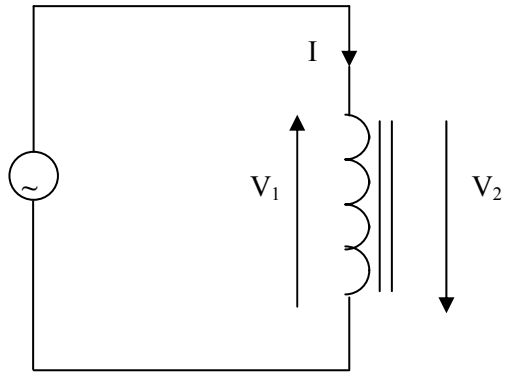
The value of an inductor indicates how much energy it can store in its magnetic field, and hence how effectively it can oppose attempts to change the current flowing through it. The value of an inductor is measured in henrys; with the abbreviation H. Typical values are measured in micro-henrys (μH) or milli-henrys (mH).

The value of an inductor depends on the number of turns and spacing between the turns of wire – the more turns, the higher the value. It also depends on the *permeability* of the material inside the coil, which may be air (for an *air-cored* inductor) or a metallic core (typically made of iron ferrite). The permeability of the core affects the strength of the magnetic field that will be caused by a current flowing through the inductor. Since iron ferrite has much higher permeability than air, a ferrite-cored inductor will have a greater inductance than an air-cored inductor with the same number of turns. Although iron ferrite cored inductors have higher inductance than air-cored inductors, they also have higher losses, especially at radio frequencies. Air core inductors may be wound with stiff wire, in which case they can be self supporting, or they may be wound on a plastic former.

Inductance is usually abbreviated "L", since "I" is already taken for current!

Inductors in A.C. Circuits

Consider the following circuit, which shows an A.C. voltage source connected to an inductor.



The effect of the A.C. supply is to continually attempt to change the current flowing through the inductor. This will change the magnetic field, which will in turn induce a voltage across the inductor that will oppose any change to the current flowing through the inductor.

For example, suppose the voltage source is attempting to increase the current flowing in the direction of *I*. Then the induced voltage will be in the direction V_1, opposing the increase in the current. However when the voltage starts trying to reduce the current flowing in the direction *I*, the induced voltage will be in the direction of V_2, now opposing the attempt to reduce the current flowing in the direction of I, and so on.

The fact that the induced voltage always opposes the change in the flow of current does not mean that one cannot change the flow of current in an inductor. It just means that the current flowing through an inductor cannot change instantaneously; it will always take some time (depending on the value of the inductance and the voltage applied) to reach the final value.

Inductive Reactance

Because in A.C. circuits the current is always changing, and inductors oppose any attempt to change the current flowing through them, inductors oppose the flow of current in an A.C. circuit. However this opposition is not resistance, since the inductor does not dissipate any power – it merely "borrows energy from" and "returns energy to" the circuit, just like a capacitor. As with capacitors, the opposition to the flow of current exhibited by an inductor in an A.C. circuit is *reactance*.

Consider the effect of frequency. The higher the frequency, the greater the rate at which the flow of current is changing. Since the inductor is effectively acting to oppose changes in the flow of current, it will exhibit higher reactance at high frequencies (where the current is changing fast) than at low frequencies (where current is changing slowly). This is evident in the formula giving the reactance of an inductor:

$$X_L = 2 \pi f L$$

Where X_L is the reactance of the inductor in ohms, π the mathematical constant pi (approximately 3,14), f the frequency in Hertz and L the inductance in Henrys. Note that there is no minus sign – the reactance of an inductor is always positive. It is also proportional to the frequency: if the frequency is doubled, the reactance is doubled, and if the frequency is halved the reactance is halved.

Ohm's Law and Reactance

Once you have determined the reactance of an inductor, you can apply Ohm's law to calculate the current or voltage in a circuit by replacing resistance with the magnitude of the reactance, |X|. For example, suppose a 1V signal at a frequency of 1 MHz (10^6 Hz) is applied across an inductance of 10 µH (10^{-5} H). The reactance of the inductor *at this frequency* can be found as follows:

$$\begin{aligned} X_L &= 2 \pi f L \\ &= 2 * 3{,}14 * 10^6 * 10^{-5} \\ &= 62{,}8\ \Omega \end{aligned}$$

The current flowing through the inductor can be calculated using ohm's law, with resistance replaced by the magnitude of the reactance:

$$\begin{aligned} I &= V / |X| \\ &= 1 / 62{,}8 \\ &= 0{,}016 \\ &= 16\ mA \end{aligned}$$

Note that although reactance is measured in ohms, it is not the same as resistance, so resistances and reactances cannot be added together.

Phase Relationship between Voltage and Current

The voltage across an inductor always *leads* the current flowing through the inductor by 90º. Conversely, the current flowing through an inductor *lags* the voltage across the inductor by 90º. The 90º phase difference between the voltage and current means that no power is dissipated by a perfect inductor – energy that is taken from the circuit and stored in the magnetic field during one part of the cycle is returned to the circuit during another part of the cycle.

Real inductors are made of electrical wire that has some resistance. Although the resistance is usually small, some power is dissipated due to the resistance of the wire.

A useful acronym to remember the phase relationship of the voltages and currents in inductors and capacitors is "CIVIL". The first three letters "CIV" mean "in a capacitor (C), current (I) leads voltage (V)". The last three letters mean "voltage (V) leads current (I) in an inductor (L)".

Inductors in Series in Parallel

Inductors in series and parallel behave similarly to resistors in series and parallel. For inductors in series,

$$L_{EQUIV} = L_1 + L_2 + \ldots$$

while for inductors in parallel,

$$1/L_{EQUIV} = 1/L_1 + 1/L_2 + \ldots$$

For example, if a 4,7 µH inductor is connected in parallel with a 3,3 µH inductor, the equivalent inductance could be found as follows:

$$\begin{aligned} 1/L_{EQUIV} &= 1/L_1 + 1/L_2 + \ldots \\ &= 1/(4{,}7 * 10^{-6}) + 1/(3{,}3 * 10^{-6}) \\ &= 212\,766 + 303\,030 \\ &= 515\,796 \end{aligned}$$

so $$\begin{aligned} L_{EQUIV} &= 1 / 515\,796 \\ &= 1{,}9\ \mu H \end{aligned}$$

Summary

Inductors store energy in their magnetic fields. As the current through an inductor changes, the changing magnetic field induces a voltage across the inductor that acts to oppose the change to the current flowing through the inductor. This is called "self inductance".

In A.C. circuits, inductors exhibit a reactance proportional to frequency. The formula for the reactance of an inductor is:

$$X_L = 2\pi f L$$

Ohm's law can be applied using the magnitude of a reactance in place of a resistance

$$V = I\,|X| \qquad \text{or} \qquad |X| = V / I \qquad \text{or} \qquad I = V / |X|$$

The voltage across an inductor *leads* the current flowing through the inductor by 90º. The phase relationships between voltage and current in capacitors and inductors can be remembered using the acronym "CIVIL".

The equivalent inductance of two or more inductors in series is given by:

$$L_{EQUIV} = L_1 + L_2 + \ldots$$

The equivalent inductance of two or more inductors in parallel is given by:

$$1/L_{EQUIV} = 1/L_1 + 1/L_2 + \ldots$$

Revision Questions

1 The characteristic back - EMF which a collapsing magnetic field causes in a coil is called:

a. Mutual inductance.
b. Self inductance.
c. Magnetic flux.
d. The solenoid effect.

2 What is the unit of inductance?

a. The henry.
b. The coulomb.
c. The farad.
d. The ohm.

3 A small air-core coil has an inductance of 5 microhenry. What do you have to do if you want a 5 millihenry coil with the same physical dimensions?

a. The coil must be wound on a non-conducting tube.
b. The coil must be wound on an iron core.
c. Both ends of the coil must be brought around to form the shape of a doughnut, or toroid.
d. The coil must be made of a heavier-gauge wire.

4 For radio frequency power applications, with which type of inductor would you get the least amount of loss?

a. Magnetic wire.
b. Iron core.
c. Air-core.
d. Slug-tuned.

5 In an inductive circuit, the alternating current produced in relation to the applied EMF is:

a. Lagging by 90 degrees.
b. 180 degrees out of phase.
c. Leading by 90 degrees.
d. In phase.

6 The phase shift between voltage and current in an inductor is:

a. 90 degrees.
b. 45 degrees.
c. 360 degrees.
d. In phase.

7 The reactance of an inductor:

a. Remains constant with changing frequency.
b. Increases with increasing frequency.
c. Decreases with increasing frequency.
d. Increases with decreasing frequency.

Chapter 10 - Tuned Circuits

Inductors and capacitors can be combined in series and parallel to form circuits that have the ability to accept or reject signals of particular frequencies. These circuits, which are called *tuned circuits*, are of great importance in radio.

Reactances in Series

Both capacitors and inductors exhibit *reactance* in A.C. circuits. The reactance depends on frequency according to the formulae:

$$X_C = -1 / (2 \pi f C)$$

and $$X_L = 2 \pi f L$$

When reactances are connected in series – for example, two capacitors or a capacitor and an inductor – then the reactances can be added to give the equivalent reactance of the two reactances in series:

$$X_{EQUIV} = X_1 + X_2 + ...$$

For example, suppose we connect two 100 pF (10^{-10} F) capacitors in series. At a frequency of 10 MHz (10^7 Hz), the reactance of each of the capacitors individually is:

$$\begin{aligned} X_C &= -1 / (2 \pi f C) \\ &= -1 / (2 * 3{,}14 * 10^7 * 10^{-10}) \\ &= -1 / 0{,}006\ 28 \\ &= -159\ \Omega \end{aligned}$$

So the equivalent reactance of the two reactances in series is:

$$\begin{aligned} X_{EQUIV} &= X_1 + X_2 \\ &= -159 + -159 \\ &= -318\ \Omega \end{aligned}$$

Of course there is another way to find this result. Since we have two capacitors of the same value (100 pF) in series, the equivalent capacitance must be half the capacitance of the individual capacitors, or 50 pF ($5 * 10^{-11}$ F). We can calculate the reactance of this equivalent 50 pF capacitance at 10 MHz (10^7 Hz) as follows:

$$\begin{aligned} X_C &= -1 / (2 \pi f C) \\ &= -1 / (2 * 3{,}14 * 10^7 * 5 * 10^{-11}) \\ &= -1 / 0{,}003\ 14 \\ &= -318\ \Omega \end{aligned}$$

Reactances in Parallel

Similarly, the formula for the equivalent reactance of two reactances in parallel is:

$$1/X_{EQUIV} = 1/X_1 + 1/X_2 + ...$$

For example, if we take our two 100 pF (10^{-10} F) capacitors, which each have a reactance of –159 Ω at 10 MHz, and connect them in parallel, then the equivalent reactance is found as follows:

$1/X_{EQUIV} = 1/X_1 + 1/X_2 + ...$
$= 1/-159 + 1/-159$
$= -0,012\ 6$

so $X_{EQUIV} = 1/-0,012\ 6$
$= -79,5\ \Omega$

Once again this makes sense since the two 100 pF capacitors connected in parallel are equivalent to a single 200 pF ($2 * 10^{-10}$ F) capacitor, with a reactance at 10 MHz of:

$X_C = -1 / (2 \pi f C)$
$= -1 / (2 * 3,14 * 10^7 * 2 * 10^{-10})$
$= -1 / 0,012\ 6$
$= -79,5\ \Omega$

The Series Tuned Circuit

Of course you might well ask, why bother to learn the formulae for reactances in series and parallel if we can calculate the same results using the formulae for capacitors and inductors in series and parallel that we already know? Good question; the answer can be found in the following circuit, which shows an inductor and a capacitor connected in series.

6,5 µH 39 pF

A Series Tuned Circuit

Suppose we want to calculate the equivalent total reactance of these two components at 10 MHz (10^7 Hz). We can't use the formula for inductors in series or the formula for capacitors in series, since the circuit contains one of each. So instead we must calculate the individual reactances of each component at a frequency of 10 MHz, and then use the formula for reactances in series.

The reactance of the inductor is found as follows:

$X_L = 2 \pi f L$
$= 2 * 3,14 * 10^7 * 6,5 * 10^{-6}$
$= 408\ \Omega$

The reactance of the capacitor is given by:

$X_C = -1 / (2 \pi f C)$
$= -1 / (2 * 3,14 * 10^7 * 39 * 10^{-12})$
$= -1 / 0,006\ 908$
$= -408\ \Omega$

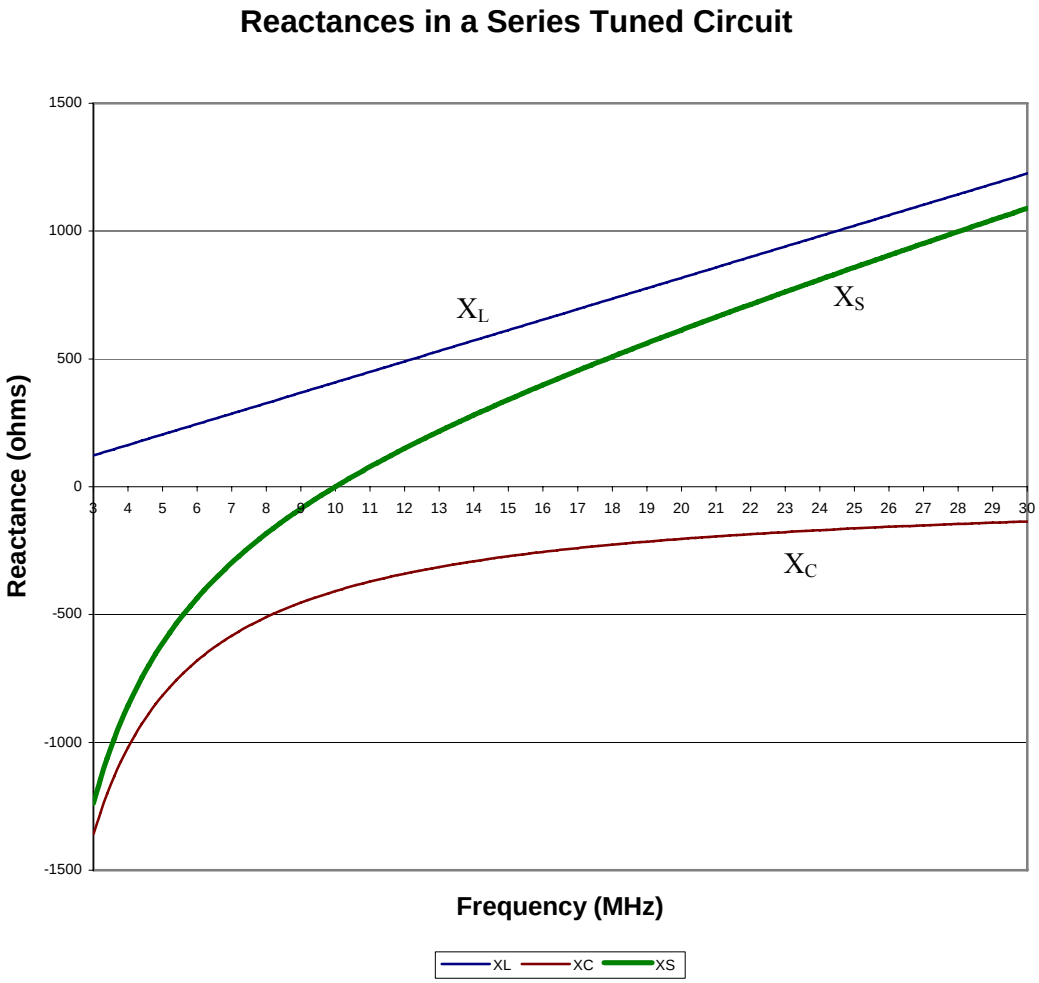
So the combined reactance of the inductor and capacitor in series at 10 MHz is

$X_{EQUIV} = X_L + X_C$
$= 408 – 408$
$= 0\ \Omega$

That's right – zero! The capacitor has reactance, and the inductor has reactance, but at this frequency (10 MHz) the positive reactance of the inductor exactly cancels out the negative

reactance of the capacitor, leaving no reactance at all! The frequency at which the positive and negative reactances cancel out is known as the *resonant frequency* of the circuit. The circuit itself is called a *series resonant* circuit or a *series tuned* circuit.

Since the reactance of the inductor *increases* with frequency, while the reactance of the capacitor *decreases* with frequency (if you forget about the minus sign), this canceling out will only happen at one specific frequency. At any other frequency, the circuit will exhibit either inductive (positive) or capacitive (negative) reactance. The graph below shows the inductive reactance X_L (which is always positive), capacitive reactance X_C (always negative) and the combined reactance of the series circuit X_S. As you can see, the combined reactance is negative (capacitive) below the resonant frequency of 10 MHz, and positive (inductive) above the resonant frequency.

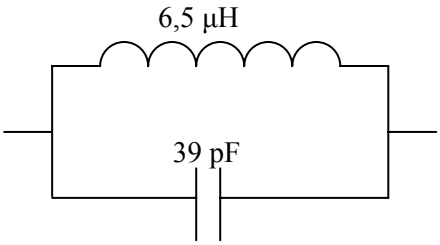


The series tuned circuit is very useful in radio electronics as the low reactance near the resonant frequency means that current can easily flow in the circuit near this frequency; while the high reactance at other frequencies will oppose the flow of current at frequencies other than the resonant frequency. In this way, a series tuned circuit can be used to accept signals with frequencies near the resonant frequency, while rejecting other signals.

The Parallel Tuned Circuit

Having seen the strange and interesting behaviour we get when we connect an inductor and capacitor in series naturally raises the question of what would happen if we were to connect

them in parallel. To save us unnecessary calculations, we may as well choose the same values – L = 6,5 μH and C = 39 pF.



A Parallel Tuned Circuit

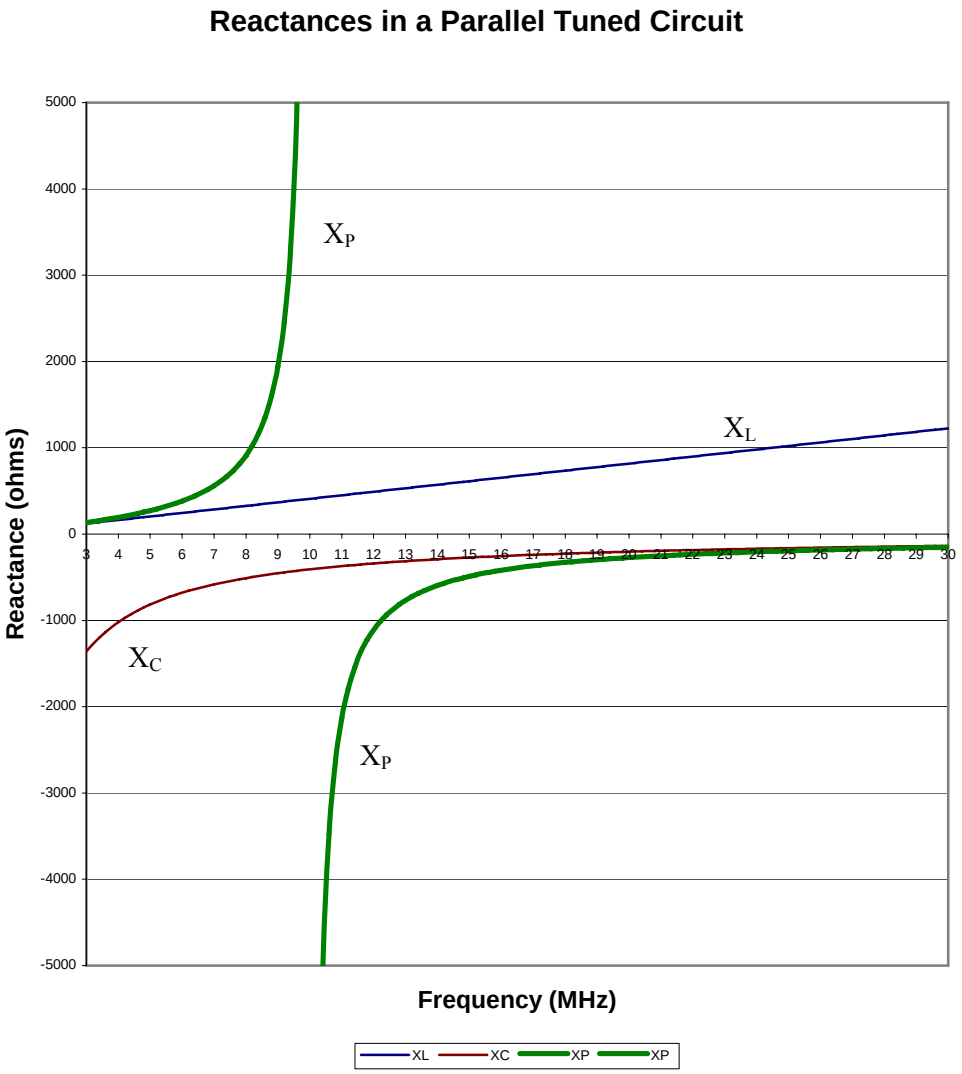
Once again we will calculate the combined reactance at 10 MHz – since this was the resonant frequency for the series tuned circuit, perhaps it will also show some interesting behaviour in this *parallel tuned circuit*.

From the formula for reactances in parallel, we know that

$$1/X_{EQUIV} = 1/X_L + 1/X_C$$
$$= 1/408 + 1/-408$$
$$= 0{,}002\ 45 - 0{,}002\ 45$$
$$= 0$$

so $X_{EQUIV} = 1 / 0$
$= ????$

What has happened here? Once again the positive inductive reactance has cancelled out the negative capacitive inductance, but this time it has left the zero in the denominator (bottom) of a fraction, which means that the result is undefined. However if we plot a graph showing the reactances for a range of frequencies, we will understand what is happening better.



Once again the inductive reactance is always positive, while the capacitive reactance is always negative. This time however the combined reactance of the tuned circuit starts slightly positive (inductive) and rapidly gets more and more positive as the resonant frequency is approached. However at the resonant frequency it instantaneously transitions from being a very high positive (inductive) reactance to being very high negative (capacitive) reactance. No wonder the exact value at resonance is undefined.

As a result, a parallel tuned circuit has a high reactance near resonance while its reactance is small away from the resonant frequency. This means that a parallel tuned circuit can be used to block signals near its resonant frequency, while allowing signals of other frequencies to pass relatively easily.

Circulating Current in a Parallel Tuned Circuit

A parallel tuned circuit has two components that are capable of storing energy. The inductor stores energy in its magnetic field; and the capacitor stores energy in the electric field between its plates. At resonance, energy is constantly being transferred from the capacitor to the inductor and back again.

As the capacitor charges up, a voltage develops between its plates. This voltage causes a current to flow through the inductor, which generates a magnetic field. As the capacitor

discharges the voltage across its plates drops, which tends to reduce the current flowing through the inductor. However an inductor will resist any attempt to change the current flowing through it. The magnetic field of the inductor collapses, inducing a potential difference into the inductor that acts to keep the current flowing in the same direction as it was before. This current flow now charges the capacitor up again, but with the opposite polarity to before. As the capacitor charges a voltage develops across its plates. This voltage causes current to flow through the inductor in the reverse direction, which generates a magnetic field, and so on.

So the parallel tuned circuit acts somewhat like a pendulum, continually transferring energy between two different forms. (In the pendulum, these forms are the potential energy when the pendulum is stationary at the top of its arc, and the kinetic energy when the pendulum is moving at maximum speed at the bottom of its arc).

One result of this is that the *circulating current* that flows in a parallel tuned circuit – that is, the current flowing around the circuit containing the capacitor and the inductor – can be much larger than the current that the parallel tuned circuit is drawing from the rest of the circuit. In practical circuits, it is not uncommon to have a circulating current that is 100 times the size of the current that the parallel tuned circuit is drawing from the external circuit.

Calculating the Resonant Frequency

We have seen that in both a *series tuned circuit* and a *parallel tuned circuit*, something interesting happens at the *resonant frequency* which is where the reactance of the capacitor and inductor have the same magnitude (value) but one is positive and the other is negative so they cancel each other out. We can derive a formula for the resonant frequency as follows:

At resonance, the magnitude of the capacitive and inductive reactances are equal, so

$$2\pi f L = 1 / (2\pi f C)$$

so $$f^2 = 1 / (4\pi^2 L C)$$

and $$f = 1 / (2\pi \sqrt{LC})$$

You do not need to know the derivation, but you should be able to apply the result. For example, let us calculate the resonant frequency of a series or parallel circuit consisting of a 6,5 µH inductor and a 39 pF capacitor:

$$f = 1 / (2\pi \sqrt{LC})$$

$$= 1 / (2 * 3{,}14 * \sqrt{6{,}5 * 10^{-6} * 39 * 10^{-12}})$$

$$= 1 / (6{,}28 * \sqrt{253{,}5 * 10^{-18}})$$

$$= 1 / (6{,}28 * 1{,}59 * 10^{-8})$$

$$= 1 / 10^{-7}$$

$$= 10^{7} \text{ Hz}$$

$$= 10 \text{ MHz}$$

Circuit Losses and the Quality Factor

The discussion so far has ignored circuit losses. For example, all practical inductors have some resistance as well as their inductance, and capacitors also have some losses although these are typically negligible compared to the losses caused by the resistance of the inductor.

The effect of these losses is that in a practical series tuned circuit, although at resonance the *reactance* would be zero, there would still be some small *resistance*. In a parallel tuned circuit, the effect of circuit losses is to limit the reactance at resonance to a high but finite value, rather than being completely undefined (or "infinite") as predicted by the maths.

The extent of circuit losses is expressed by a number called the "Quality Factor", or "Q Factor" or just the "Q" of the tuned circuit. A high Quality Factor means low circuit losses, while a low Qualify Factor means high circuit losses. The quality factor is defined as the reactance of either the inductor or the capacitor at resonance divided by the circuit resistance. So

$$Q = X_L / R$$
$$= -X_C / R$$

(The minus sign in the second line is just to take account of the fact that capacitive reactance is itself negative and ensure that we come up with a positive Q). The Quality Factor of practical tuned circuits is typically between 50 and 200.

The Quality Factor is related to two other properties of the tuned circuit:

1. The ratio of circulating current in a parallel tuned circuit to the current drawn by the tuned circuit is the same as the Q factor. So in a parallel tuned circuit with a Q of 100, the circulating current will be 100 times greater than the current drawn from the rest of the circuit.

2. The selectivity of the circuit – that is, its ability to allow desired signals through while blocking undesired signals. The greater the Q of the tuned the circuit, the greater its selectivity.

Summary

The series tuned circuit has a low reactance near its resonant frequency, and a high reactance at other frequencies. Series tuned circuits are often used to allow signals near the resonant frequency to pass, while blocking signals at other frequencies.

The parallel tuned circuit has a high reactance near its resonant frequency, and a low reactance at other frequencies. Parallel tuned circuits are often used to block signals near the resonant frequency, while allowing signals at other frequencies to pass.

The resonant frequency of a series or parallel tuned circuit may be calculated as

$$f = 1 / (2 \pi \sqrt{LC})$$

The Quality Factor ("circuit Q") is defined as the reactance of either the inductor or the capacitor at resonance divided by the circuit resistance. A tuned circuit with a high Q is more selective than a tuned circuit with a low Q.

The circulating current in a parallel tuned circuit may be many times the current drawn by the tuned circuit. The ratio between the circulating current and the current flowing into the tuned circuit is the same as the Quality Factor.

Revision Questions

1 At one particular frequency, resonance of a capacitor and inductor takes place. At this frequency:
a. Inductive reactance is nil.
b. Capacitive reactance is nil.
c. The impedance is nil.
d. The capacitive and inductive reactances are equal.

2 The parallel tuned circuit impedance at resonance is:
a. Low.
b. High.
c. Infinitely high.
d. Equal to 10.

3 The series tuned circuit impedance at resonance is:
a. Low.
b. High.
c. Infinitely high.
d. Equal to 10.

4 The Q of a resonant circuit determines the:
a. Losses of the circuit.
b. Value of the capacitance required for resonance.
c. The inductor value required for resonance.
d. Value of increased current through the coil and capacitor at resonance.

5 The selectivity of a resonant circuit is greater if the Q factor:
a. Is low.
b. Decreases to 1.
c. Is high.
d. Remains low.

6 The resonant frequency of a tuned circuit consisting of a 10 nF capacitor in parallel with a 10 µH inductor is approximately:
a. 500 kHz.
b. 5 MHz.
c. 50 MHz.
e. 500 MHz.

7 You have a 100 µH inductor and wish to create a tuned circuit with a resonant frequency of 3,500 MHz. What value of capacitor would you require?
a. 2,1 pF.
b. 12 pF.
c. 21 pF.
d. 120 pF.

8. You have a 10 pF capacitor and wish to create a tuned circuit with a resonant frequency of 10 MHz. What value of inductor do you require?
a. 2,5 µH.
b. 10 µH.
c. 25 µH.
d. 100 µH.

Chapter 11 - Decibel Notation

In amateur radio we often deal with ratios of powers. For example, the *gain* of an amplifier is the ratio of its output power to its input power. These ratios can be very large or very small. For example, the gain of a typical amateur radio receiver – the ratio between the output power into the speaker or headphones to the input power from the antenna – is in the region of 100 000 000 000 000. That's an amplification of a hundred trillion times! While we could use scientific notation to represent these large numbers (the one above is 10^{14}), another way of expressing the ratio of two powers is commonly used. This is the "decibel".

The unit "bel" was first used by telephone engineers at Bell Laboratories (now AT&T) and was named after Alexander Graham Bell (1847-1922), the inventor of the telephone and founder of Bell Laboratories. The "decibel" is simply one tenth of a bel, which turned out to be a more popular size. One decibel represents roughly the minimum discernable change in the loudness of an audio signal. The abbreviation for the decibel is "dB", which is also often used in general conversation such as "your signal is S9 plus 20 dB".

A ratio of two powers can be expressed in decibels as follows:

$$dB = 10 \log_{10} (PR)$$

where PR is the ratio of two powers (e.g. $PR = P_1 / P_2$), "dB" is the same ratio expressed in decibels, and "$\log_{10}$" means the mathematical logarithm to the base 10. If you are not familiar with logarithms then don't panic – once we have explored a couple of the properties of decibels we will see that there is a simple way to calculate many common values.

Adding Decibels

A fundamental property of decibels is that when two ratios expressed in decibels are added, it is equivalent to multiplying the original ratios. For example, a ratio of 2 times is 3 dB and a ratio of 10 times is 10 dB. If we add the decibel representations we get 3 dB + 10 dB = 13 dB, which is equivalent to a ratio of 20 times. This is the same as we get if we multiply the ratios: 2 * 10 = 20. This bit of magic is possible because of the use of the logarithm function in the definition of the decibel.

Example

In a radio receiver the radio frequency (RF) amplifier has a gain of 6 dB; the intermediate frequency (I.F.) amplifier has a gain of 110 dB and the audio frequency (A.F.) amplifier has a gain of 20 dB. What is the total gain of the receiver?

If the gains of the amplifiers had been expressed as simple ratios (P_{OUT}/P_{IN}) then we would have to *multiply* the ratios together to get the total gain. However since the gains are expressed in decibels, we can *add* them to get the total gain. So in this case the total gain is 6 dB + 110 dB + 20 dB = 136 dB.

Representing Losses

The decibel can also be used to represent losses, i.e. situations where a signal gets smaller. If you calculate the decibel equivalent of a ratio that is less than 1, then the formula gives a *negative* number. For example we can calculate the decibel equivalent of a power ratio of 0,1 as follows:

$$dB = 10 \log_{10} (PR)$$

$= 10 \log_{10} (0,1)$
$= 10 * -1$
$= -10$ dB

So, for example, an attenuator that reduces a signal to one-tenth its original power could be described as having a *gain* of –10 dB. Note that the minus sign indicates that it is actually making the signal smaller even though it is expressed as a "gain". The same attenuator could also be described as having a *loss* of 10 dB. This time there is no minus sign because it is being described as a *loss*.

However if you add decibels together (which as we have seen is equivalent to multiplying the original ratios), then you should express all the ratios as either gains or losses before adding them together. You can't add a decibel representing a *gain* to one representing a *loss*.

Example

An attenuator with a loss of 6 dB is added before the RF amplifier in a receiver. Before adding the attenuator, the receiver had a gain of 136 dB. What is the total gain of the receiver with the attenuator?

Because we can't add the 6 dB *loss* of the attenuator to the 136 dB *gain* of the receiver, we first convert express the attenuator's *gain* as –6 dB. Then we can calculate the total gain of the receiver by adding the –6 dB gain of the attenuator to the 136 dB gain of the receiver to get the answer 130 dB.

Finally, a gain of exactly 1 (i.e. a signal that gets neither stronger nor weaker) can be represented as 0 dB. This makes sense, since *adding* 0 dB to a ratio represented in decibels will not change it; just as *multiplying* a ratio by 1 won't change it either.

Quick and Easy Decibel Conversions

Some commonly used ratios are easily converted to decibels. These are shown in the table below:

Power Ratio	Decibels	Power Ratio	Decibels
1000 000	60 dB	0,000 001	-60 dB
100 000	50 dB	0,000 01	-50 dB
10 000	40 dB	0,000 1	-40 dB
1 000	30 dB	0,001	-30 dB
100	20 dB	0,01	-20 dB
10	10 dB	0,1	-10 dB
5	7 dB	0,2	-7 dB
4	6 dB	0,25	-6 dB
2	3 dB	0,5	-3 dB
1	0 dB		

You don't need to remember all the powers of ten (the numbers 10, 100, 1 000 etc). If a ratio consists of a 1 followed by any number of zeros, then it to convert it to decibels simply multiply the number of zeros by ten. For example, 1 000 000 has 6 zeros so it is equivalent to 60 dB (the number of zeros times ten).

Using these values it is possible to easily calculate the decibel representation of many other common ratios. For example, what is the decibel equivalent of a ratio of 20? Well 20 is not in

the table, but 2 and 10 are, and 20 = 2 * 10. However we know that multiplying ratios is the same as adding their decibel equivalents, so the decibel equivalent of 20 must be the decibel equivalent of 2 plus the decibel equivalent of 10. So the answer is 3 dB + 10 dB = 13 dB, which is the decibel equivalent of 20.

Of course this works the other way round as well. Suppose we want to calculate the ratio represented by 27 dB. Although 27 dB is not in the table, we know that 27 dB = 20 dB + 7 dB, and both the values *are* in the table. Since adding decibels is equivalent to multiplying ratios, the ratio represented by 27 dB is the ratio represented by 20 dB *multiplied by* the ratio represented by 7 dB. So the answer is 100 * 5 = 500, which is the ratio represented by 27 dB.

Expressing Voltage Ratios as Decibels

Throughout this module We have stressed that decibel notation is used to express the ratio of two *powers*. However because there is a relationship between voltage and power, decibels are also sometimes used to express the ratio between two *voltages*. Now the relationship between voltage and power can be expressed as

$$P = V^2 / R$$

Because power is proportional to the voltage *squared*, if the voltage is doubled then the power will be multiplied by 4; if the voltage is increased by a factor of 10 then the power will be multiplied by 100. (Note that this does not depend on the resistance, it will hold true for any resistance as long as the same resistance is used to calculate the power before and after the voltage is increased.)
Because of this, a modified formula is used to express a ratio of voltages in decibels:

$$dB = 20 \log_{10} (VR)$$

where *VR* is the ratio of two voltages and *dB* is the same ratio expressed in decibels. Note that the constant "10" in the formula used for power ratios is replaced by "20" in the formula for voltage ratios. This is to take into account the V^2 factor in the formula for power. In other words, when we representing a voltage ratio in decibels, we are still representing a ratio between two powers. In this case, however, it is the notional power that would be dissipated by some (unknown) load if the voltages in question were applied across the load.

If you want to express a voltage ratio in decibels using the "quick and easy" conversions outlined above, then you should *square* the voltage ratio (multiply it by itself) to convert the voltage ratio to a power ratio before converting it into decibels.

Note that expressing voltage ratios as decibels is a confusing and potentially misleading exercise. Wherever possible, deal with *power* ratios not voltage ratios.

For example, suppose the input voltage of an amplifier is 10 μV and the output voltage is 1 mV. The input and output resistances of the amplifier are both 50 Ω and we want to calculate the gain of the amplifier in decibels.

The input and output powers can be found from

$$P_{IN} = V^2 / R = (10 * 10^{-6})^2 / 50 = 2 \text{ pW}$$

$$P_{OUT} = V^2 / R = (10^{-3})^2 / 50$$

= 20 nW

Having calculated the powers, we can express them as a ratio and then convert it to decibels:

P_{OUT}/P_{IN} = 20 nW / 2 pW
= 10 000
= 40 dB

An alternative way to reach the same answer would be to take the voltage ratio

VR = 1 mV / 10 µV
= 100

Then square this to find the power ratio

PR = 100^2
= 10 000

And then convert this express this as 40 dB. (Remember, the number of zeros multiplied by ten!) However this alternative approach will only work if the input and output resistances are equal. The first method – calculating the actual input and output powers – will work whatever the input and output resistances, as long as you know what they are.

Expressing Power Levels in dBW and dBm

In the new Radio Regulations, the power levels that apply to amateur transmissions are not expressed in watts as before, but rather in dBW. The unit dBW means "decibels referenced to 1 W". It is a way to express actual powers in decibel notation. Note that one cannot express an actual power – say 100 W – in decibels since decibels are used to express the *ratio* of two powers. However if you make one of the two powers a standard reference level, then by expressing the ratio of the other power to this standard reference level you can communicate an actual power level. One of the common reference levels is 1 W, and the resulting unit is given the abbreviation "dBW". For example, the maximum power level specified for a Class A1 (ZS) license is 26 dBW. This means "26 dB over 1 W". Since 26 dB is a ratio of 400, 26 dBW means 400 W.

A related unit is decibels over 1 mW. This unit is abbreviated "dBm". For example, the sensitivity of most amateur receivers is around –130 dBm, meaning "130 dB less than 1 mW". (The minus sign means that the level is *less than* the reference level of 1 mW). This is equivalent to the incredibly small value of 10^{-16} W, or 0,1 femto-watts!

Summary

The decibel is a logarithmic unit used to express the ratio of two powers. The ratio of two powers can be converted to decibels using the formula

$$dB = 10 \log_{10} (PR)$$

Adding two ratios expressed in decibels is equivalent to multiplying the original ratios. However both of the figures added must express either a gain or a loss; you cannot add a gain to a loss. To convert a gain to a loss or vice-versa, simply put a minus sign before it. If a ratio consists of a 1 followed by any number of zeros, then it to convert it to decibels simply multiply the number of zeros by ten.
A ratio of two voltages can be expressed in decibels using the formula

$$dB = 20 \log_{10} (VR)$$

This is equivalent to first converting the voltage ratio to a power ratio by squaring it, and then expressing the resulting power ratio in decibels. This will only give the correct result if both voltages are applied across the same resistance.

Although absolute powers cannot be expressed in decibels, they can be expressed in dBW (decibels referenced to 1 W) or dBm (decibels referenced to 1 mW).

Revision Questions

1 An increase in power from 0,25 W to 1,25 W is equal to a gain of:

a. 3 dB.
b. 7 dB.
c. 10 dB.
d. 1 dB.

2 A transmitter has a power output of 100 W. This is connected to an antenna with 11 dB gain by means of a coax cable with a loss of 1 dB. The ERP (effective radiated power) of the transmitter, coax and antenna combined is:

a. 11 W.
b. 111 W.
c. 1 000 W.
d. 2 000 W.

3 A -20 dB attenuator is placed in line with a 40 V RMS signal. Assuming the impedances all remain constant what will the reduced signal level be?

a. 2 V.
b. 10 V.
c. 20 V.
d. 4 V.

4 A power gain of 4 is equivalent to:

a. 3 dB.
b. 6 dB.
c. 10 dB.
d. 16 dB.

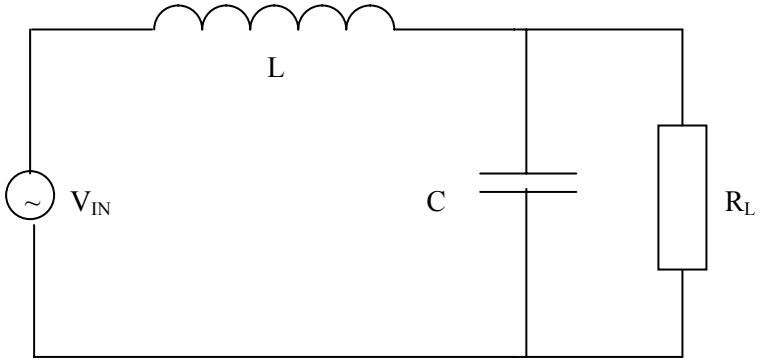
5 A signal with a power of 1 mW is applied to the input of an amplifier that has a gain of 13 dB. The power of the output signal will be:

a. 5 mW.
b. 10 mW.
c. 20 mW.
d. 100 mW.

Chapter 12 - Filters

Filters are electrical circuits that allow signals of particular frequencies to pass, while blocking signals of other frequencies. They can be used, for example, to select the signal that a radio receiver is tuned to, while blocking the signals that it is not tuned to.

The Low-Pass Filter



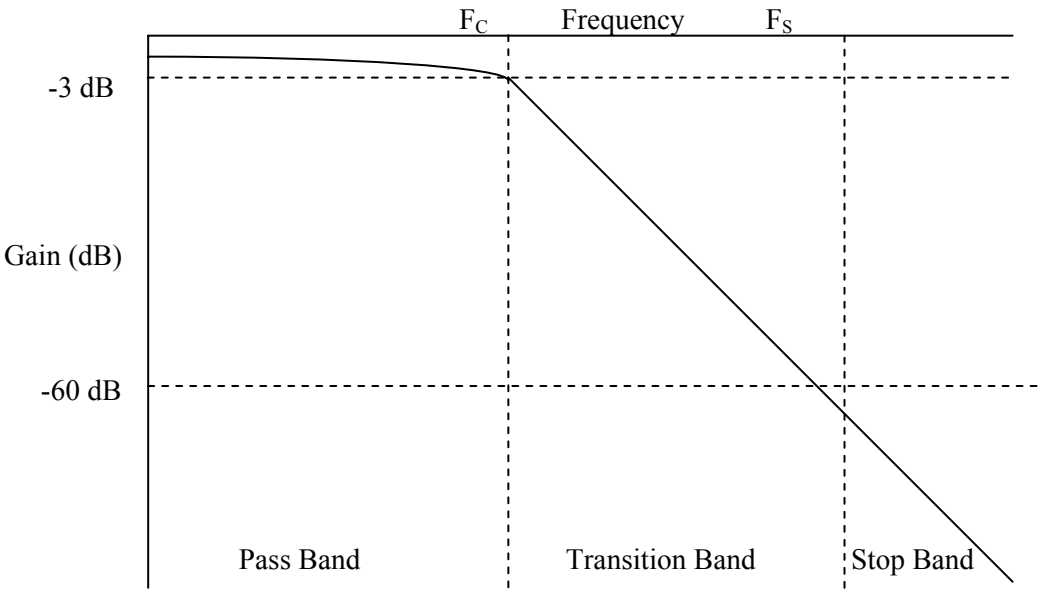
An input voltage V_{IN} is applied across a voltage divider consisting of an inductor L and a capacitor C in parallel with a resistive load, R_L.

Although we are not in a position to analyze this circuit quantitatively, we can get a good qualitative idea of what happens. When the frequency of the input voltage is low, the inductor has low reactance while the capacitor has high (negative) reactance. This means there is little opposition to current flowing through L, but significant opposition to current flowing through C. As a result, most of the input voltage is applied across the load resistance R_L, and power is efficiently transferred to the load.

Now consider what happens when the frequency is high. Since the reactance of an inductor is proportional to the frequency, L will have high reactance. On the other hand, the reactance of a capacitor decreases with frequency, to C will have a low impedance. This means that the inductor provides significant opposition to the flow of current; and what current is able to flow is mostly diverted through the capacitor rather than flowing through the load. As a result, little power is transferred to the load.

This circuit is called a “low-pass filter” because it allows low frequency signals to pass (in other words, to be efficiently coupled to the load), while blocking high frequency signals.

A graph can be plotted showing the *frequency response* of the filter – that is, its gain at different frequencies.



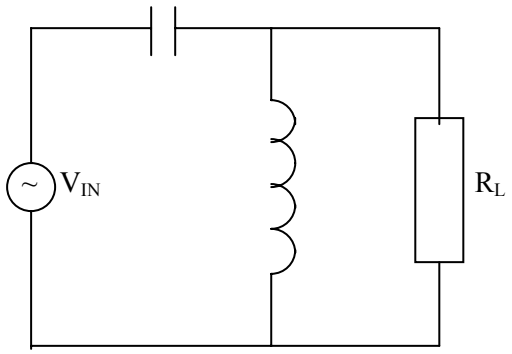
The Frequency Response of a Low-Pass Filter

The *cutoff frequency* F_C is the frequency at which the attenuation of the filter is 3 dB (i.e. the gain is –3 dB). At this frequency, half the input power reaches the load. For a low-pass filter, signals with frequencies lower than the cut-off frequency have relatively little attenuation; these signals are in the *pass band* of the filter.

Signals with frequencies higher than F_S are greatly attenuated – in this case by 60 dB or more. These signals are in the *stop band* of the filter. Signals with frequencies between F_C and F_S are somewhat attenuated. These frequencies are sometimes called the *transition band* of the filter since it is in transition between the pass band and the stop band.

Most amateur radio transmitters have a low-pass filter after the final power amplifier to attenuate any *harmonics* of the output frequency. Harmonics are multiples of the output frequency caused by distortion in the amplifier, so for example a transmitter that is transmitting on a frequency of 3,5 MHz might have harmonics on 7 MHz, 10,5 MHz, 14 MHz, 17,5 MHz, 21 MHz and so on. It is very difficult to design a power amplifier that does not generate any harmonics, and in any case such an amplifier would probably be quite inefficient. However it is easy to use a low-pass filter at the output to pass the desired frequencies and attenuate the harmonics to an acceptably low level.

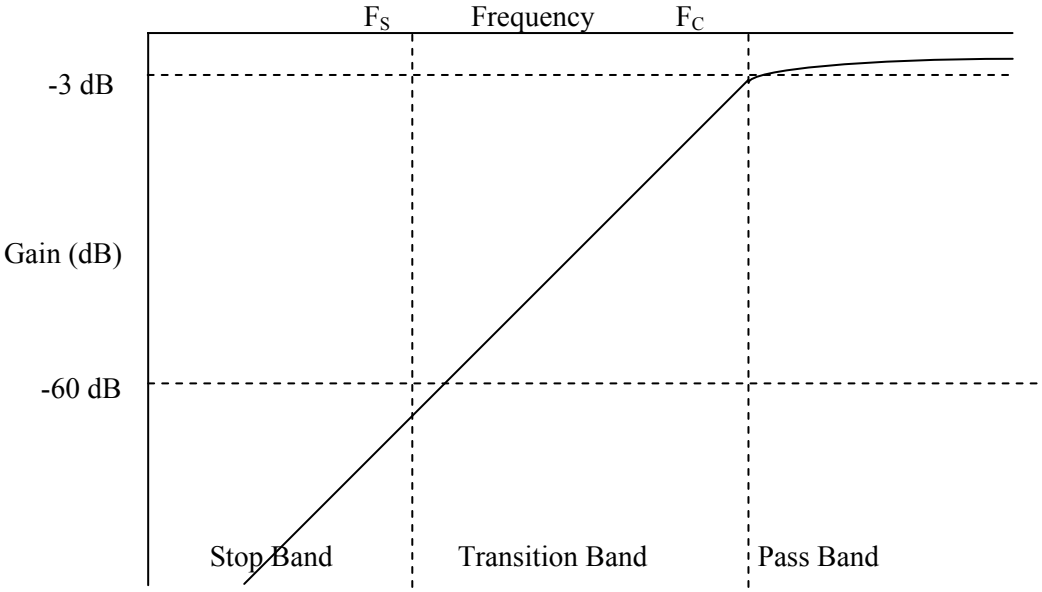
The High-Pass Filter



Once again the input voltage V_{IN} is applied to a voltage divider, but this time the capacitor and inductor in the voltage divider have been swapped. At low frequencies, the capacitor has

high reactance and so opposes the flow of current; while the inductor has low reactance so the current that does flow is diverted through the inductor rather than flowing through the load.

At high frequencies, the capacitor has low reactance, so does little to oppose the flow of current. The inductor has high reactance, so most of the current flows through the load resistor R_L rather than through the inductor. This circuit is called a "high-pass" filter because it allows high frequency signals to pass (in other words to be efficiently coupled to the load) while blocking low frequency signals. The frequency response of a high-pass filter looks something like this:



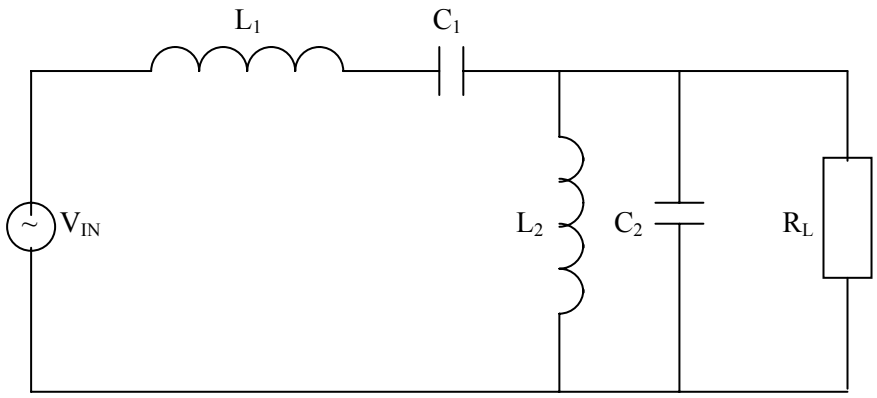
The Frequency Response of a High-Pass Filter

Once again, the cutoff frequency is the frequency at which the attenuation of the filter is 3 dB (the *half-power* point), while We have chosen to measure the stop-band from the point where the attenuation is 60 dB.

High-pass filters are often used in the input stages of receivers to reject the very strong radio signals found in the medium wave broadcast band from 500 kHz to 1,5 MHz so they do not overload the receiver, while allowing signals in the amateur bands starting at 1,8 MHz to pass.

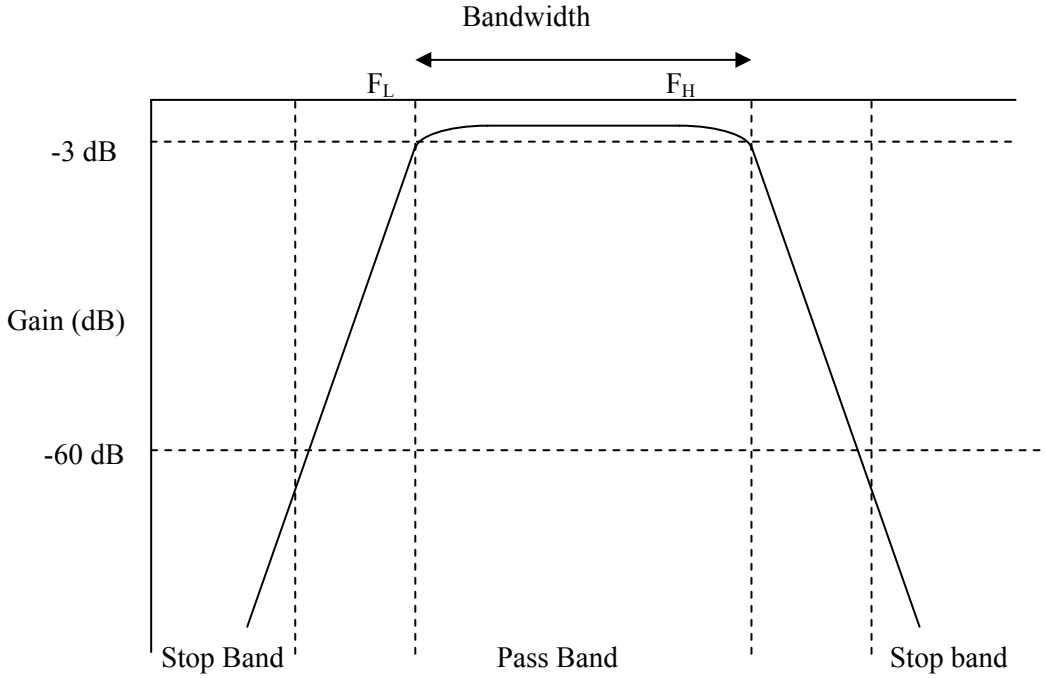
The Band-Pass Filter

Band-pass filters pass signals in a certain frequency range known as the *pass band* and reject signals with frequencies above or below the pass band. They can be constructed using series and parallel tuned circuits. For example, consider the circuit below:



Once again we have a circuit resembling a voltage divider, although this time it is made up of two tuned circuits – a series tuned circuit consisting of L_1 and C_1 in series with the source, and a parallel tuned circuit consisting of L_2 and C_2 across the load. Assume that the two tuned circuits have the same resonant frequency. Near this frequency, the series tuned circuit has low reactance while the parallel tuned circuit has very high reactance, so almost the entire input voltage appears across the load. This is the pass band of the filter.

However at frequencies well above or below the resonant frequency, the series tuned circuit has a high impedance while the parallel tuned circuit has a low impedance, so very little of the input voltage appears across the load. This is the stop band of the filter.



The Frequency Response of a Band-Pass Filter

The band-pass filter has two cutoff frequencies, a high cut-off labeled F_H and a low cut-off labeled F_L. Both cutoff frequencies are measured at the point where the output from the filter is 3 dB below the input to the filter (the *half-power* points). The *bandwidth* of the filter is the difference (in Hertz) between the high cutoff frequency and the low cutoff frequency. For example, if the high cutoff frequency is 2 700 Hz and the low cutoff frequency is 300 Hz then the bandwidth is 2 700 – 300 = 2 400 Hz. The *centre frequency* of a band-pass filer is the frequency half way between the high cutoff frequency and the low cutoff frequency; in this case it would be 1 500 Hz.

Most amateur receivers use band-pass filters to allow signals from a particular amateur band to enter the receiver while rejecting signals from other amateur bands. This is called a *preselector*.

Crystal Filters

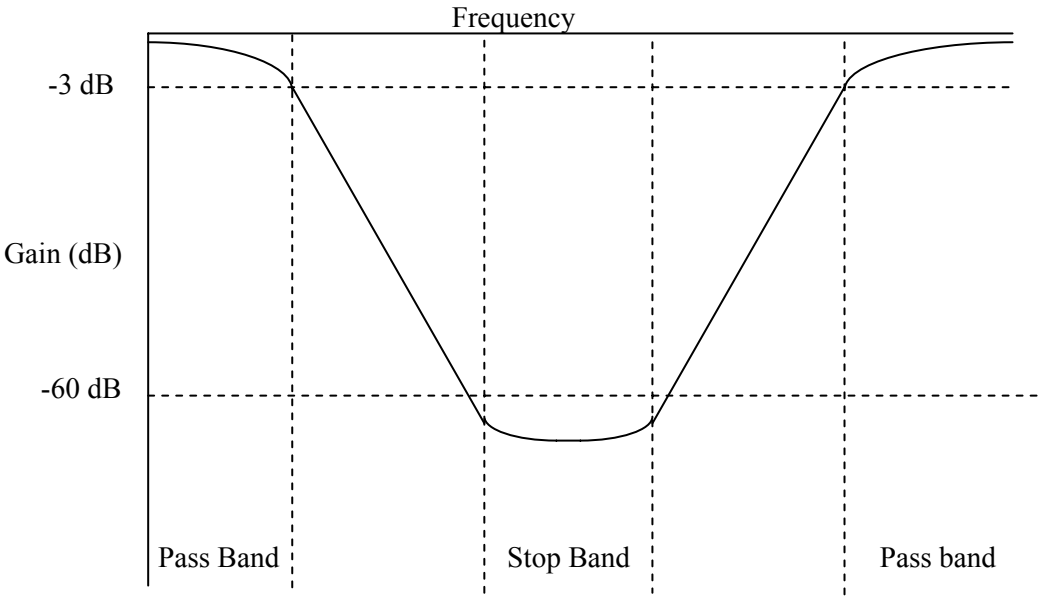
Band-pass filters can also be implemented using quartz crystals. These have a piezoelectric property, which means that a voltage applied to the crystal causes a slight physical movement of the crystal; and physical movements of the crystal will in turn cause a voltage to appear across it. Quartz crystals have very similar properties to tuned circuits and can be used to make highly selective band-pass filters. These "crystal filters" are responsible for the selectivity – that is, the ability to distinguish one signal from another – of many modern amateur receivers and transceivers.

Although crystal filters are very selective – that is, their bandwidth is very narrow in comparison with the centre frequency of the filter – they have the disadvantage that they only work at a single fixed frequency. That is, a crystal filter cannot be tuned to different frequencies. When we look at the design of superhet receivers we will see how this limitation is overcome while allowing the receiver to take advantage of the exceptionally good selectivity of crystal filters.

Amateur receivers and transceivers often allow you to select different bandwidth crystal filters for different purposes. Some of the common bandwidths are 2,4 kHz for normal phone (SSB) operation, 1,8 kHz for phone operation under difficult conditions (often used in contests) and between 250 Hz and 500 Hz for CW (Morse Code) operation. Most transceivers come with one or two basic filters (for example, just a 2,4 kHz filter) but additional filters can often be purchased, although they can be quite expensive.

The Band-Stop Filter

A band-stop filter works in the opposite way to a band-pass filter. Frequencies in a certain range (the stop-band) are attenuated, while frequencies either above or below those frequencies are passed. Amateur receivers and transceivers often provide a manually adjustable band-stop filter that can be used to attenuate undesired signals, for example a carrier generated by someone tuning up close to the frequency that you are listening to. These are known as "notch filters" because they allow you to "notch out" undesired signals.



The Frequency Response of a Band-Stop Filter

Summary

Low-pass filters allow signals with frequencies below the cut-off frequency to pass with little attenuation, while significantly attenuating signals with frequencies well above the cut-off frequency. High-pass filters allow signals with frequencies above the cut-off frequency to pass with little attenuation, while significantly attenuating signals with frequencies well below the cut-off frequency. In both cases, the cut-off frequency is measured from the point where the signal is attenuated by 3 dB; this is also known as the "half power" point.

Band-pass filters allow signals with frequencies between the low and high cut-off frequencies to pass, while attenuating signals with frequencies significantly higher or lower than the pass-band. The bandwidth of a band-pass filter is the difference between the high cut-off and low cut-off frequencies. Crystal filters are highly selective band-pass filters. Band-stop filters attenuate signals with frequencies in a particular range, while allowing signals outside that frequency range to pass.

Revision Questions

1 A band pass filter:
a. Allows all frequencies to pass.
b. Attenuates all frequencies.
c. Allows signals between two frequencies to pass.
d. increases bandwidth of a receiver.

2 A band stop filter :
a. Allows all frequencies to pass.
b. Attenuates all frequencies.
c. decreases bandwidth of a receiver.
d. Attenuates signals between two frequencies.

3 A Low pass filter:
a. Attenuates all signals above a known cut-off frequency.
b. Introduces harmonics.
c. Removes RF signals from an input signal.
d. Requires the use of high gain amplifiers.

4 A high pass filter:
a. Introduces harmonics.
b. Removes RF signals from an input signal.
c. Requires the use of high gain amplifiers.
d. Attenuates all signals below a known cut-off frequency.

5 What is a circuit called which passes electrical energy above a certain frequency, but blocks electrical energy below that frequency?
a. An input filter.
b. A low-pass filter.
c. A high-pass filter.
d. A band-pass filter.

6 The purpose of a low pass filter is to:
a. attenuate all frequencies apart from a specific one.
b. pass all frequencies apart from a specific one.
c. pass all signals below a specified frequency but attenuate frequencies above it.
d. attenuate all signals below a specified frequency but pass frequencies above it.

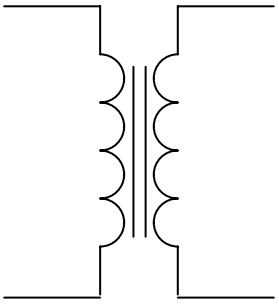
7 The purpose of a high pass filter is to:
a. attenuate all frequencies apart from a specific one.
b. pass all frequencies apart from a specific one.
c. pass all signals below a specified frequency but attenuate frequencies above it.
d. attenuate all signals below a specified frequency but pass frequencies above it.

Chapter 13 - The Transformer

Theory of Operation

The transformer is used to change (transform) the voltage and current of an AC signal. It consists of two or more windings wound on a common former. One of these windings is called the *primary winding*. The rest of the windings are known as *secondary windings*.

In operation, an A.C. voltage is applied to the primary winding. This generates a fluctuating magnetic field, which induces a voltage into the secondary windings. This property is called *mutual inductance* to distinguish it from the *self inductance* that is characteristic of inductors. The circuit symbol for a transformer is shown below:



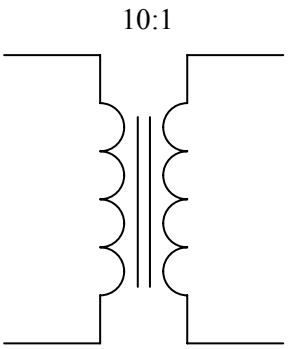
So which is the primary winding and which is the secondary? This is not shown in the circuit symbol, but must be deduced from the context. The primary winding is whichever winding has power applied to it.

Turns Ratio

The turns ratio of a transformer specifies the relative number of turns on the primary and on the secondary. The number of turns on the primary is always specified first. For example, a 5:1 transformer has five times as many turns on the primary as on the secondary; a 1:3 transformer has three times as many turns on the secondary as on the primary.

Note that the turns ratio does not specify the actual number of turns, just the ratio of turns on the primary with respect to the number of turns on the secondary. For example, a transformer with 200 turns on the primary and 20 turns on the secondary would be described as a 10:1 transformer because there are ten times as many turns on the primary as on the secondary.

The turns ratio is often shown in numbers on the circuit symbol, for example:



Voltage Ratio

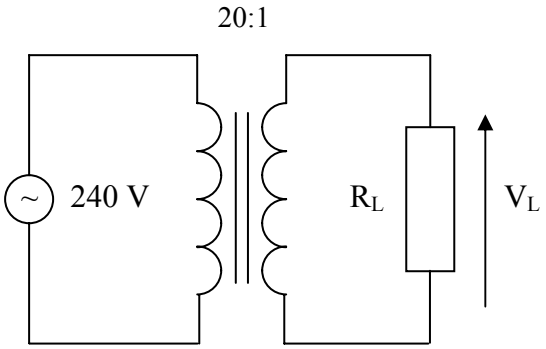
The voltages across the different windings of a transformer follow a very simple rule known as the *transformer principle*:

The Transformer Principle: *The ratio of the voltage on the primary winding to the voltage on the secondary winding is the same as the ratio of the number of turns on the primary winding to the number of turns on the secondary winding (the turns ratio)*.

This can be written mathematically as follows:

$$V_P/V_S = N_P/N_S$$
so $$V_S = V_P N_S / N_P$$

Where N_P is the number of turns on the primary and N_S the number of turns on the secondary. For example, consider the circuit below:



240 V A.C. is applied across the primary of a 20:1 transformer. What is the voltage V_L across the load? From the transformer principle, the ratio of the voltage on the primary to the voltage on the secondary must be the same as the ratio of the number of turns on the primary to the number of turns on the secondary, which is 20:1. So if the voltage on the 240 V, then the voltage on the primary must be one twentieth of this, or 12 V. Alternatively you can use the formula to get the same result:

$$V_S = V_P N_S / N_P$$
$$= 240 * 1 / 20$$
$$= 12\ V$$

A transformer with more turns on the secondary than on the primary will have a higher voltage across the secondary than across the primary; this is called a *step-up* transformer because it "steps the primary voltage up" to a higher secondary voltage. A transformer with fewer turns on the secondary than on the primary will have a lower voltage across the secondary than across the primary and is called a *step-down* transformer. For example, a 1:5 transformer is a step-up transformer, while a 10:1 transformer is a step-down transformer.

Current Ratio

For a transformer with a single secondary winding, the ratio of the current in the secondary to the current in the primary is the same as the ratio of the number of turns in the primary to the number of turns in the secondary (the turns ratio). Note that this is the opposite way round to the voltage ratio, so a step-up transformer (which has a greater voltage across the secondary than across the primary) will have a smaller current flowing in the secondary than in the primary; while a step-down transformer (which has a smaller voltage across the secondary

than across the primary) will have a larger current flowing in the secondary than in the primary.

Mathematically this can be expressed as follows:

$$I_S/I_P = N_P/N_S$$
or $$I_S = I_P N_P / N_S$$

So for the example above, suppose the current flowing in the primary is 1 A. Then the current in the secondary can be found from

$$I_S = I_P N_P / N_S$$
$$= 1 * 20 / 1$$
$$= 20 A$$

So the transformer has converted a high voltage and low current to a low voltage and high current. It is interesting to compare the power supplied by the transformer's secondary to the power drawn by the primary.

The power drawn by the primary is:

$$P_P = V_P I_P$$
$$= 240 * 1$$
$$= 240 W$$

The power supplied to the load by the secondary is:

$$P_S = V_S I_S$$
$$= 12 * 20$$
$$= 240 W$$

Since the power supplied to the load by the secondary is identical to the power drawn by the primary, the transformer has not dissipated any power. In practical transformers there is usually some small power dissipation caused by the resistance of the windings and eddy currents flowing in the transformer core.

The ability of transformers to efficiently and simply convert high voltages to low voltages and vice-versa is the principle reason why the mains supply in all countries is A.C. since it allows the very high voltages used in the distribution network (which minimize the I^2R heating losses in the power distribution cables) to be efficiently converted to lower voltages like 240 V for domestic use.

Impedance Ratio

We haven't quite finished with our example circuit yet. Since we know the voltage across the load resistance and the current flowing through the load, we can calculate the resistance:

$$R_L = V_S / I_S$$
$$= 12 / 20$$
$$= 0,6 \Omega$$

We can also calculate the resistance that the primary winding appears to have to the voltage source driving it.

$$R_P = V_P / I_P$$

$= 240 / 1$
$= 240\ \Omega$

Note that this "resistance" is not the actual resistance of the primary winding, which would typically have a resistance of less than 1 Ω. It is rather an apparent resistance caused by the fact that power is being drawn from the primary winding. In this case though the power is ending up in the secondary circuit and is not being dissipated by the transformer itself but rather by the load.

Another way to look at it is that the voltage source driving the primary is "seeing" the load resistance, but the transformer has transformed the actual value of the resistance, just as it has transformed the voltage and current values. We can derive a general rule for this resistance transformation:

$$\begin{aligned} R_L &= V_S / I_S \\ &= (V_P N_S / N_P) / (I_P N_P / N_S) \\ &= (V_P / I_P) (N_S^2 / N_P^2) \\ &= R_P N_S^2 / N_P^2 \\ &= R_P (N_S / N_P)^2 \end{aligned}$$

conversely,

$$R_P = R_L (N_P / N_S)^2$$

Where R_P is the "apparent resistance" of the primary winding. The fact that the load resistance in the secondary circuit causes a different (but related) resistance to appear in the primary circuit is known as "impedance transformation". Impedance is a general concept that combines resistance and reactance, which is covered in more detail in another module.

These equations tell us that the resistance in the primary circuit and the resistance in the secondary circuit are related by the *square* of the turns ratio. The resistance transformation works in the "same direction" as the voltage transformation so for a step-down transformer, where the voltage across the secondary is smaller than the voltage across the primary, the resistance in the secondary circuit will also be smaller than the resistance in the primary circuit. Similarly, a step-up transformer will "step up" the resistance in the primary circuit to a larger resistance in the secondary circuit. However note that the amount by which impedances are transformed is not the same as the amount by which voltages are transformed since the impedance transformation depends on the *square* of the turns ratio, while the voltage transformation depends on the turns ratio (not squared).

For example, suppose you have an audio amplifier designed to drive a 200 Ω load but you want to connect it to an 8 Ω speaker instead. You could use a transformer to convert the impedances. You need a 200:8 or 25:1 impedance transformation from the primary (connected to the amplifier output) to the secondary (connected to the speaker). This would be accomplished using a 5:1 transformer. (Note that the turns ratio, 5:1, is the *square root* of the impedance ratio, 25:1.)

Applications

Transformers are widely used in amateur radio. The most obvious example is in power supplies, where the mains voltage of 240 V must be transformed to a voltage suitable for running radio equipment, typically 12 V.

Transformers are also widely used for *impedance matching* within transmitter and receiver circuits. For example, an antenna system typically has an impedance of 50 Ω, while the RF

amplifier of a typical receiver would generally have a higher input impedance. Since maximum power is transferred when the source and load impedances are equal, a transformer might be used to match the impedance of the antenna to the input impedance of the RF amplifier.

Summary

An A.C. voltage applied to the *primary winding* of a transformer generates a fluctuating magnetic field that induces a voltage in the *secondary winding*. This is known a *mutual inductance*.

The ratio of the voltage on the primary winding to the voltage on the secondary winding is the same as the ratio of the number of turns on the primary winding to the number of turns on the secondary winding (the *turns ratio*).

$$V_S = V_P N_S / N_P$$

A transformer with more turns on the secondary than on the primary is a *step-up* transformer; one with fewer turns on the secondary than on the primary is a *step-down* transformer.

The ratio of the current in the secondary to the current in the primary is the inverse of this, so

$$I_S = I_P N_P / N_S$$

The overall effect is that the power in the primary circuit is equal to the power in the secondary circuit, so a perfect transformer does not dissipate power.

By transforming voltages and currents, transformers also transform the load resistance to an apparent resistance in the primary winding. The transformation occurs in the same "direction" as the voltage transformation, but according to the *square* of the turns ratio:

$$R_P = R_L (N_P / N_S)^2$$

Revision Questions

1 The principle of operation of a transformer is based upon:

a. Static electricity.
b. Potential difference.
c. Electrostatics.
d. Electromagnetic induction.

2 Transformers transfer energy from one coil to another by means of:

a. Inductive coupling.
b. Static discharge.
c. Capacitance.
d. Electrical conduction.

3 A transformer with a turns ratio of 1:8 is called:

a. a step down transformer.
b. a step up transformer.
c. low current transformer.
d. a high-tension transformer.

4 A transformer nameplate shows a figure of 1:4. If 12 V AC is applied to the primary winding, what is the voltage on the secondary terminals?

a. 3 V.
b. 48 V.
c. 16 V.
d. 8 V.

5 On which electrical principle is the functioning of a transformer based?

a. Stray capacitance.
b. Mutual inductance.
c. Eddy currents.
d. Circuit resistance.

6 What is the turns ratio of a transformer to match an audio amplifier having an output impedance of 200 Ω to a speaker having a load impedance of 10 Ω?

a. 4,47 to 1.
b. 14,14 to 1.
c. 20 to 1.
d. 400 to 1.

7 The operating principle of a transformer may be described as:

a. A varying magnetic field intersecting a conductor and creating a potential difference.
b. A varying electric field intersecting a conductor and creating a potential difference.
c. A varying current in a conductor setting up a static magnetic field.
d. A varying voltage in a conductor setting up a static magnetic field.

8 An impedance-matching transformer has a turns ratio of 10:1. If a 500 Ω microphone is connected to the winding with the lesser turns, it would correctly operate into a load of:

a. 5 Ω.
b. 50 Ω.
c. 50 kΩ.
d. 500 kΩ.

9 A transformer has 1 200 turns on its primary and 30 turns on its secondary. If it is connected to the mains supply (240 V), the secondary voltage will be:

a. 9 600 V.
b. 240 V.
c. 30 V.
d. 6 V.

10 A 5:1 transformer has a current of 1 A flowing in the primary winding. The current flowing in the secondary winding is:

a. 40 mA.
b. 200 mA.
c. 5 A.
d. 25 A.

Chapter 14 - Semiconductors and the Diode

Semiconductors

Semiconductors are materials where the outer electrons are more tightly bound to the nucleus than in the case of conductors, but less tightly bound than in the case of insulators. Normally at room temperature these materials behave as insulators; but when heated, the additional energy of the electrons allows the outer ones to break away from the nucleus, so allowing a current to flow if an electric potential is applied.

The semiconductors most commonly used in electronic devices are silicon and germanium. Silicon atoms have 14 electrons, arranged in three layers. The first (inner) layer has 2 electrons; the second layer has 8 electrons; and the outer layer has 4 electrons. It is the electrons in this outer layer that are only moderately bound to the nucleus. Silicon atoms normally form a crystal lattice with other silicon atoms, where each atom shares one of its outer electrons with another atom in the lattice.

When silicon is used to manufacture electronic components, it is first refined to make it very pure silicon, and then small quantities of another material are introduced. This process is known as “doping”.

N-Type Semiconductors

Suppose a very small quantity of a material with 5 electrons in its outer shell such as phosphorous or arsenic is added to very pure molten silicon and then the mixture is allowed to cool and crystallize. The silicon atoms will take up their normal crystal structure, with each atom sharing one electron with each of its four neighbouring atoms. The occasional phosphorous or arsenic atom will be forced to fit in with this structure, so it will also share one of its outer electrons with each of its four neighbouring silicon atoms. However since phosphorous and arsenic have *five* outer electrons, this will leave one electron that is not bound into the crystal structure. This electron will be free to migrate around the crystal lattice, and can serve as a charge carrier, allowing a current to flow if a potential is applied. So silicon doped in this way becomes an electrical conductor at room temperature, due to the free charge carriers. Since the charge carriers are negatively charged electrons, this is called an “N-type” semiconductor, where the N stands for “negative”.

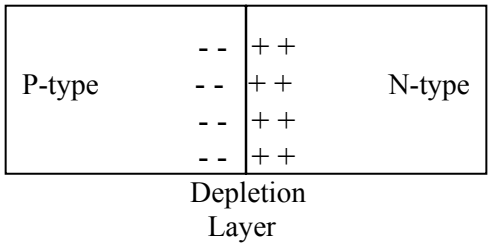
P-Type Semiconductors

Now suppose that we add a small quantity of a material with only 3 outer electrons, such as boron or aluminium, to pure molten silicon and allow it to cool and form a crystal lattice. Again the silicon atoms will take up their normal crystal structure, with each atom sharing one electron with each of its four neighbours. The occasional boron or aluminium atom will be forced to fit into the structure, sharing one of its outer electrons with each of its four neighbouring silicon atoms. However since boron and aluminium only have *three* outer electrons, one of its neighbours will have to do without a shared electron. This leaves a “hole” in the crystal lattice, a place where there ought to be an electron but isn’t one. If an electric potential is applied across such a material, the electrons will be attracted by the positive terminal and repelled by the negative terminal. However most of the electrons will be unable to move as they are rigidly bound into the lattice structure. However the electron that is on the negative potential side of the “hole” is able to move – it can move to the place in the lattice where the electron is missing, leaving it’s own place in the lattice structure empty. Another electron can fill this empty slot, leaving its place empty, and so on. In this way the “hole” can migrate across the lattice, even though it is actually electrons that are moving. Because the hole is the absence of a negatively charged electron, it behaves as though it was a positive charge free to move around the lattice, For example, if a potential difference is applied, then holes will migrate from the positive terminal to the negative terminal. In this way, holes can

also act as charge carriers, and since they behave like positively charged charge carriers, semiconductors doped in this way are known as "P-type" semiconductors.

The Junction Diode

Now suppose that a small piece of P-type semiconductor is brought into contact with a small piece of N-type semiconductor. This is called a "P-N junction".



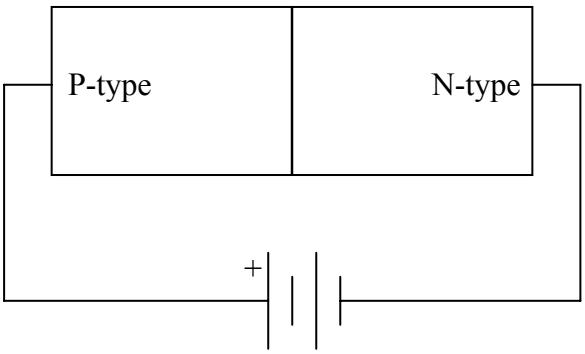
The P-N Junction

Because the N-type material has some free electrons, some of these can migrate across the boundary and "fill" some of the holes in the lattice structure of the P-type material. This will leave the P-type material negatively charged, while the N-type material will become positively charged. The process will stop when the potential difference between the P-type and N-type materials is sufficient to prevent any further electron movement across the boundary.

(Remember that the P and N type materials were *not* positively and negatively charged to start with. They were both neutral since the "extra" electrons in the N-type material were balanced by the additional positive charges on the nuclei of the phosphorous or arsenic atoms used for doping. Similarly, the "lack" of electrons in the P-type material is precisely balanced by the smaller positive charge on the nuclei of the boron or aluminum atoms.)

This forms a very thin layer called the *depletion layer* or *depletion region* at the junction between the P-type and N-type materials where there are no (or very few) free charge carriers since the free electrons from the N-type material have all been "used up" filling the holes on the P-type side of the junction. The depletion layer is usually very thin, about 0,001 mm or so.

Now suppose we apply a potential difference across the junction, with the positive terminal attached to the P-type material and the negative terminal to the N-type material.



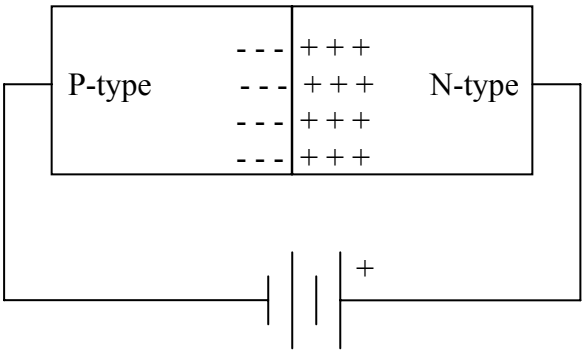
A Forward Biased P-N Junction

The battery will effectively "pump" electrons into the N-Type material and towards the depletion region. Similarly, it the positive potential applied to the P-type material will attract

electrons away from the depletion region. The net effect is that there now are charge carriers in the depletion region – electrons in the N-Type material, and holes in the P-type material – and so a current can flow. Another way of looking at it is that the depletion layer has been neutralized by the application of a potential difference across the forward biased junction.

In order to make this current flow, there must be sufficient potential difference to overcome the potential difference that existed across the junction between the N-type and P-type materials due to the migration of electrons from the N-type material to the P-type material when the depletion layer was formed. This voltage is known as the "forward bias voltage" and is typically between 0,5 V and 0,8 V in silicon P-N junctions and around 0,1 to 0,2 V for germanium junctions. If the potential difference applied is less than this then electrons won't be forced into the depletion region on the N-type side, or removed from it on the P-type side, so the depletion region will still not have any free charge carriers and no current will flow.

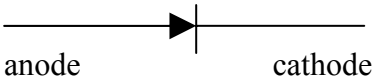
Now let's see what happens if we connect the potential the other way around, with the positive terminal attached to the N-type material and the negative terminal to the P-type material.



A Reverse Biased P-N Junction

The effect of the applied potential is to force more electrons into the P-type material, making it more negatively charged, and remove electrons from the N-type material, making it more positively charged. This simply results in more holes in the P-type material being "filled" by the additional electrons, and more of the free electrons being removed from the N-type material, increasing the depth of the depletion layer and increasing the potential difference across the junction until it equals the potential difference applied to the junction. So after a very brief initial current, no further current ill flow except a tiny current known as the "reverse leakage" current which is typically less than 1 μA. This is known a "reverse biased" junction.

The device we have described from a physical point of view is called the *junction diode*, often just "diode" for short. Its circuit symbol is shown below:

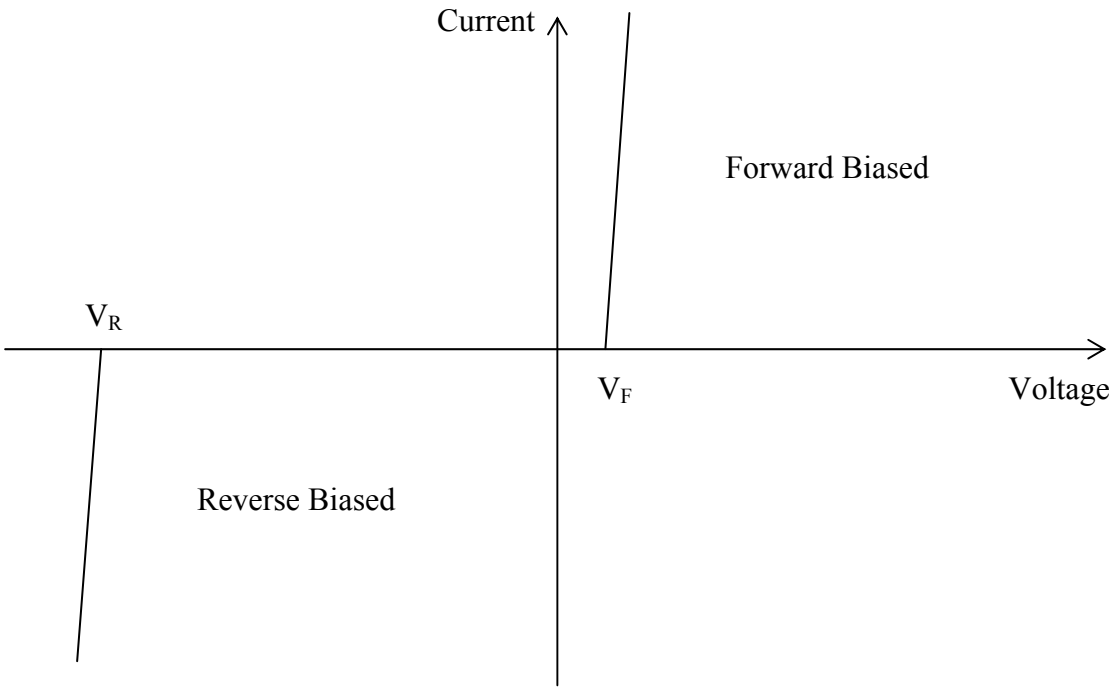


The two terminals of the diode are called the "anode" and the "cathode", and it will allow a current to flow if the anode is more positive than the cathode by 0,5-0,8 V, the forward bias voltage. This current flows in the direction of the arrow, i.e. from left to right in the diagram above.

If the diode is reverse biased then only a tiny current, the reverse leakage current, will flow. Of course if you apply a high enough potential across a reverse-biased junction then

eventually the depletion layer will break down, allowing a current to flow However in most diodes (with the exception of the *Zener* diode as we will see below) this is likely to cause permanent damage to the device.

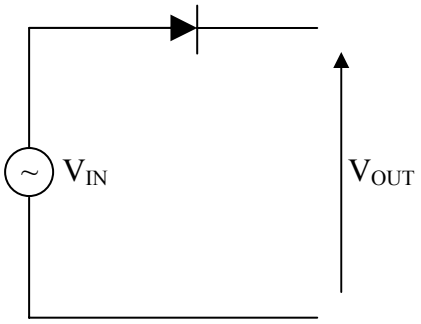
The graph below plots the voltage across the diode against the current flowing through the diode, using the convention that a positive voltage means the diode is forward biased, and a negative voltage means it is reverse biased.



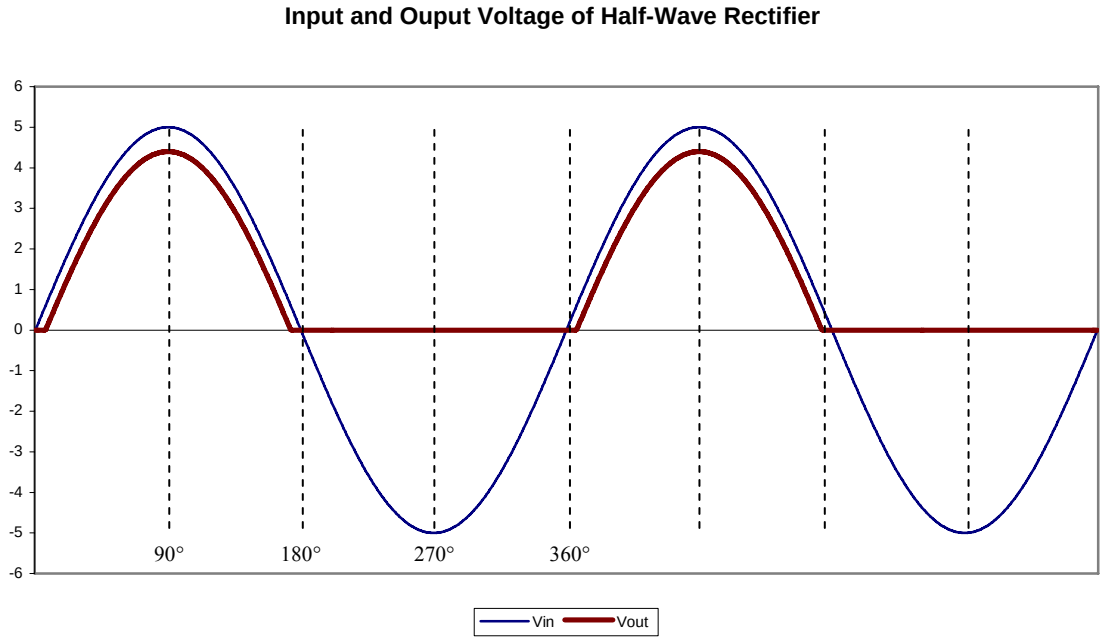
When the diode is forward biased, no current flows until the voltage applied exceeds the forward biased voltage of the diode, V_F. Once the voltage exceeds V_F, the current rises rapidly with little change in the voltage across the diode. For this reason, it is a good approximation to assume that *the voltage across a forward biased diode is always V_F, the forward bias voltage*.

The Half-Wave Rectifier

Diodes are commonly used as *rectifiers* to convert A.C. voltages into D.C. voltages. For example, consider the circuit below, which is known as a *half-wave rectifier*.

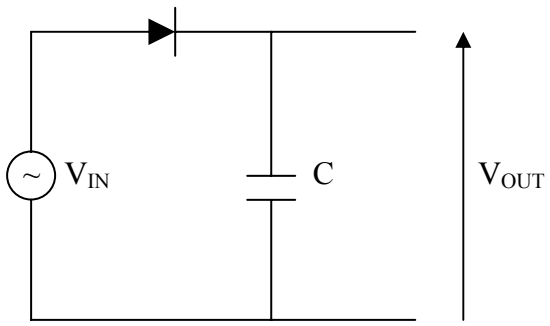


The graph below shows the input and output voltages for a half-wave rectifier



The output voltage follows the input voltage during the positive half cycles, but is always slightly less than the input voltage due to the forward-bias voltage (in this case 0,6 V) across the diode. During the negative half cycles the diode does not conduct and the output voltage is zero.

Of course the resulting voltage can hardly be considered "D.C." since it is still varying periodically. However this can be solved using a *smoothing capacitor*.

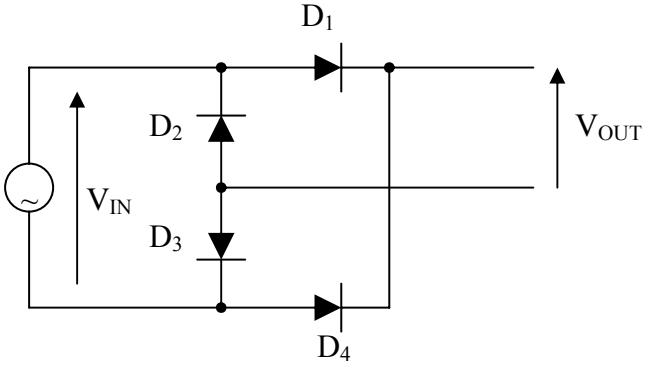


The "smoothing capacitor" C will charge up during the positive half cycle, and then discharge into the load (which is not shown) during the negative half cycle, keeping a relatively constant D.C. voltage across the load. However some traces of the original alternating voltage will remain superimposed on the D.C. output voltage; this is known as "ripple" and can cause problems such as hum in audio amplifiers. For a half-wave rectifier, the ripple has the same frequency of the A.C. signal that is rectified.

Another way of thinking of this circuit is that the capacitor C forms a simple one-element low-pass filter. The half-wave rectified sine wave contains both a D.C. component and an A.C. component, and the capacitor attenuates the A.C. ripple component while passing the D.C. output.

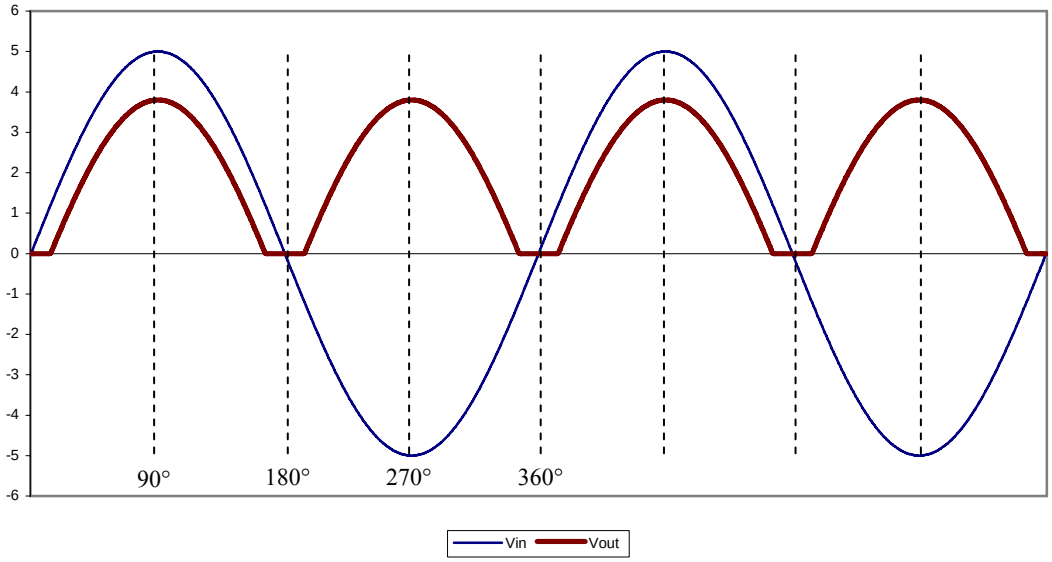
The Full-Wave Rectifier

Half-wave rectifiers suffer from excessive ripple because they give zero voltage output for just over half of each cycle. The full-wave rectifier shown below is better in this regard.



When V_{IN} is positive, D_1 and D_3 will conduct and V_{OUT} will be positive. When V_{IN} is negative, D_2 and D_4 will conduct, and V_{OUT} will still be positive! This circuit is known as a *full-wave bridge rectifier* and the graph below shows the input and output voltages.

Input and Output Voltage of Full-Wave Rectifier



The output voltage is now positive during both the positive and the negative half-cycle of the input voltage. Note that the output voltage is always 1,2 V lower than the input voltage; this is because there are now *two* 0,6 V forward voltage drops across the two conducting diodes. Note also that the frequency of the "ripple" on the output is now *twice* the input frequency.

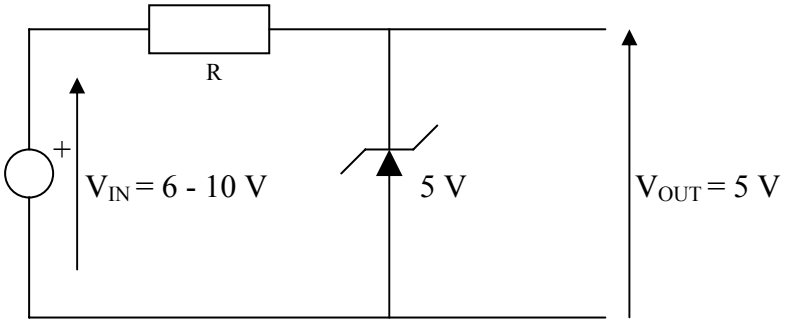
Once again, the amount of ripple on the output can be greatly reduced by using a suitable smoothing capacitor as a low-pass filter on the output.

The Zener Diode

In most diodes the reverse breakdown voltage V_R should never be exceeded or the diode may be permanently damaged. However a specific type of diode, the *Zener diode* is manufactured in which the reverse breakdown voltage can be safely exceeded. Zener diodes are also manufactured in such a way that the reverse breakdown voltage (also known as the *Zener*

voltage, abbreviated as V_Z) is specified and carefully controlled. Zener diodes are available with Zener voltages ranging between 2 V and 100 V.

Zener diodes are commonly used as voltage regulators. For example, consider the circuit below:



This shows an input voltage in the range 6 V to 10 V being applied across a *reverse-biased* Zener diode with a Zener voltage of 5 V through a resistor. The purpose of the resistor is to limit the current flowing through the Zener diode and prevent it from being destroyed. Note the circuit symbol for the Zener diode.

If the output voltage rises above 5 V, then the Zener diode conducts strongly even though it is reverse biased. The current flowing through the Zener diode causes a voltage drop across the resistor, reducing the output voltage until it is close to 5 V. In this way the circuit will maintain a fairly constant output voltage of around 5 V irrespective of both the input voltage and the current drawn by the load. In this way the reverse-biased Zener diode functions as a *voltage regulator*, maintaining a constant output voltage despite fluctuations in the input voltage and load current.

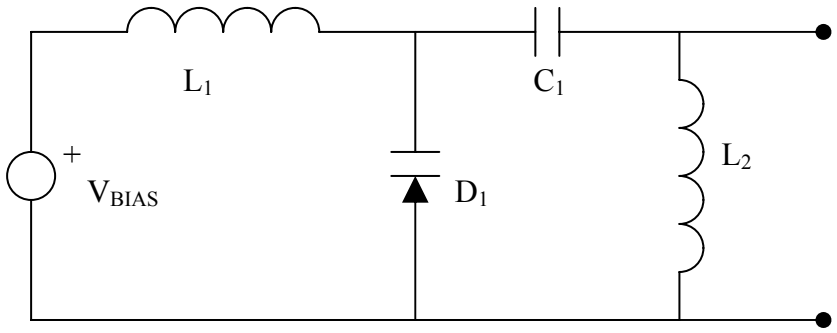
The Varicap Diode

Think back to the physical description of the reverse-biased P-N junction. It consists of two conducting materials (P-type and N-type semiconductors) separated by a thin non-conducting layer (the depletion layer).

This is very similar to the description of a capacitor: two conducting plates separated by a thin insulating layer. Indeed, all reverse-biased diodes do act as capacitors to some extent. Most diodes are designed to minimise the capacitance effect so that it becomes negligible except at very high frequencies.

However some diodes are designed to maximise this capacitance effect and to allow it to be controlled by the reverse-bias voltage applied to the diode. Remember that the greater the reverse-bias voltage, the larger the depletion layer. This is equivalent to moving the plates of a capacitor further apart – in other words, the capacitance will be reduced. These are called varicap ("variable capacitance") diodes.

A typical circuit is shown below:

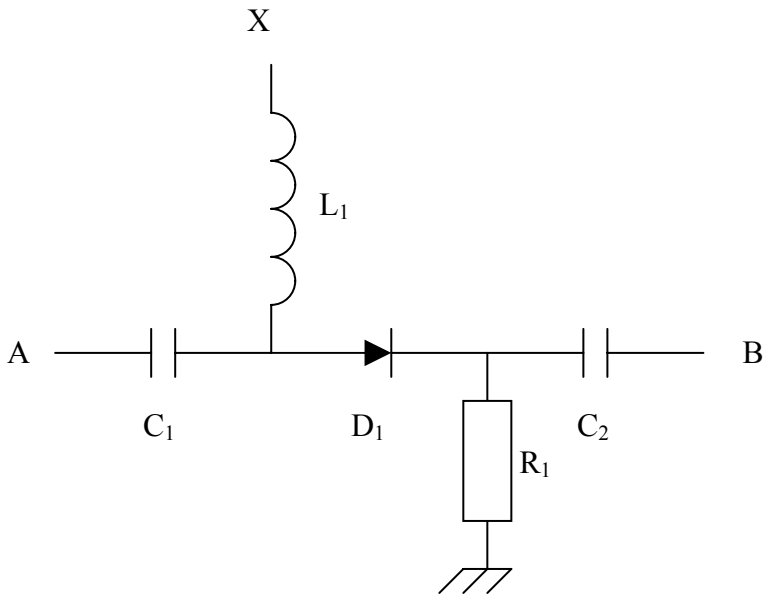


As you can see, the circuit symbol for a varicap diode is appropriately a combination of the symbols for a diode and for a capacitor.

In this circuit a D.C. bias voltage V_{BIAS} is used to reverse-bias varicap diode D_1. As the bias voltage is changed, the capacitance of D_1 is changed. This will vary the resonant frequency of the parallel tuned circuit made up of L_2, C_1 and D_1. Note that the capacitance in the parallel tuned circuit consists of C_1 in series with the varicap diode. L_1 prevents the A.C. signals present in the tuned circuit from flowing through the bias voltage source, which would introduce undesirable losses into the tuned circuit, reducing its Q. Similarly C_1 prevents the D.C. bias voltage from being shorted through L_2. Circuits like this are often use to vary the frequency of oscillators (that is, circuits which generate A.C. signals) in response to a D.C. tuning voltage.

Diodes as Switches

Although it may appear surprising at first, diodes are often used as switches in radio frequency (RF) applications.



Suppose an A.C. voltage is applied at point A. will it reach B? Well let us first assume that the control voltage X is positive with respect to the chassis. The D.C. current will flow through inductor L_1 and diode D_1, and to the chassis via resistor R_1. (The symbol below the resistor represents the conducting metallic chassis.) Because the diode is forward biased, it offers very little resistance to the flow of current, so a signal applied at point A will be coupled through capacitor C_1, diode D_1 and capacitor C_2 to point B. As long as the signal applied at A is significantly smaller than the control voltage applied at X, the signal will *not*

be rectified since the diode will be forward biased even during the negative peaks of the signal.

Now consider what happens if a negative voltage (with respect to the chassis) is applied to X. Now D_1 is reverse-biased, and will not conduct current. Again assuming that the signal applied at A is significantly smaller than the control voltage, the diode will remain reverse biased even at the positive peaks of the signal. Hence the signal will effectively be blocked by the reverse-biased diode.

The role of L_1 is simply to have high reactance at the signal frequency, preventing the signal from being coupled into the control circuitry, while allowing the D.C. control current to flow unimpeded.

Diodes are widely used in this way in modern microprocessor-controlled transceivers since the microprocessor can easily generate suitable control voltages, which can be used to perform various RF switching tasks, such as switching between different filters. The best diodes for switching tasks are PIN diodes, which have a low resistance when forward biased, and high resistance when reverse biased, combined with low capacitance. Significant capacitance across the reverse-biased diode would be disastrous in this circuit, as it would allow the signal through even when the switch was turned "off"!

Summary

Pure semiconductors do not conduct current at room temperature. However by introducing small amounts of impurities they can be made to conduct a current. The charge carriers in N-type semiconductors are negatively charged electrons; while in P-type semiconductors they are positively charged holes.

When N-type and P-type semiconductors are brought into contact, a thin *depletion layer* with no free charge carriers forms at the junction. If the junction is *forward biased* by making the P-type material positive with respect to the N-type layer then a current will flow provided the potential exceeds the forward bias voltage of the junction, which is typically around 0,6 V for silicon devices. If the junction is *reverse biased* by making the P-type material negative with respect to the N-type material then no current will flow until the *reverse breakdown* voltage is reached.

The P-N junction is used to make an electronic device called the *junction diode*. The diode has two terminals: the anode, which is connected internally to the P-type material; and the cathode, which is connected to the N-type material. If the anode is made positive with respect to the cathode by 0,6 V or so (the forward bias voltage) then a current will flow through the diode.

Diodes are commonly used as rectifiers. In a *half-wave rectifier*, only the positive half cycle is rectified, so there is a substantial amount of *ripple* at the A.C. frequency, which must be removed using a *smoothing capacitor*. In a *full-wave rectifier* both the positive and the negative half-cycles are rectified, resulting in less ripple. In a full-wave rectifier the ripple frequency is *twice* the frequency of the A.C. input.

Zener diodes have a well-controlled and specified reverse breakdown voltage, also known as the *Zener voltage*, and are designed not to be damaged by reverse breakdown. When reverse biased, Zener diodes act as voltage regulators. *Varicap diodes* exhibit a capacitance when reverse biased that decreases as the reverse bias voltage is increased. *PIN diodes* are often used as switches at radio frequencies.

Revision Questions

1 The forward biased diode junction will:
a. Allow current to flow through the junction.
b. Block current from flowing.
c. Have a high resistance.
d. Exhibit a zener diode function.

2 In a silicon diode the voltage of 0,6 V refers to:
a. The breakdown voltage.
b. The forward bias voltage.
c. The zener voltage.
d. The cut-off voltage.

3 Which of the following components is intended only to operate in the reverse biased condition?
a. A rectifier diode.
b. A zener diode.
c. A polarized capacitor.
d. A resistor.

4 The term Vz is commonly used to describe:
a. The zener diode regulating voltage.
b. The zener diode impedance.
c. The forward voltage applied to the diode.
d. The peak voltage of the rectified waveform.

5 By only allowing Alternating current to flow in one direction a diode can be used as:
a. An attenuator.
b. An amplifier.
c. A rectifier.
d. A fuse.

6 Zener diodes are used to:
a. Start oscillations.
b. Divert signals.
c. Detect modulation.
d. Regulate a DC voltage supply.

7 What function does a full-wave bridge circuit serve?
a. Amplification.
b. Coupling.
c. Rectification.
d. Isolation.

8 A circuit which only allows half of an AC waveform to pass through is called:
a. A regulator.
b. A bridge circuit.
c. An attenuator.
d. A half wave rectifier.

9 A four-diode circuit to produce full-wave rectified DC from a transformer is called:

a. A balanced circuit.
b. A bridge rectifier.
c. A dummy load.
d. A regulator.

10 The area of a diode junction where no free holes or electrons exist is called the:

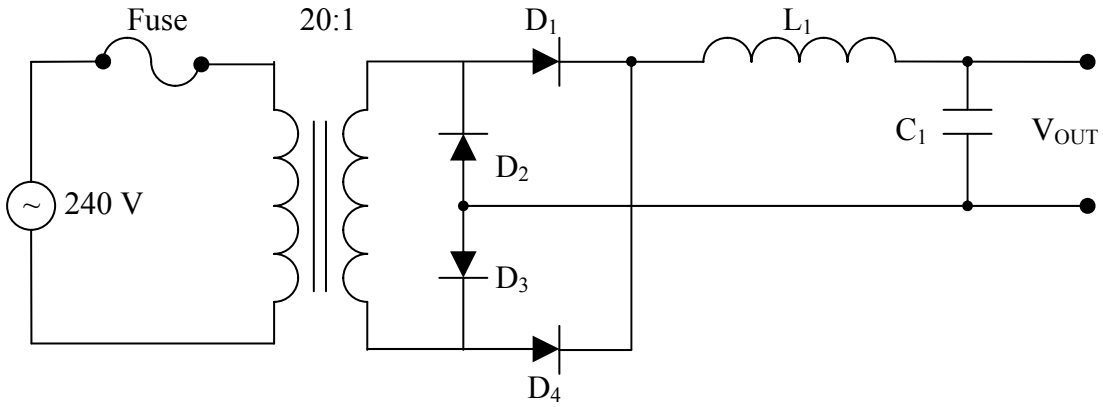
a. Anode.
b. Cathode.
c. Depletion region.
d. Semiconductor.

11 An PN type semiconductor refers to a:

a. Two pin transistor.
b. A capacitor.
c. A diode.
d. A power resistor.

Chapter 15 - The Power Supply

The circuit below shows a simple unregulated D.C. power supply that is easy to construct.



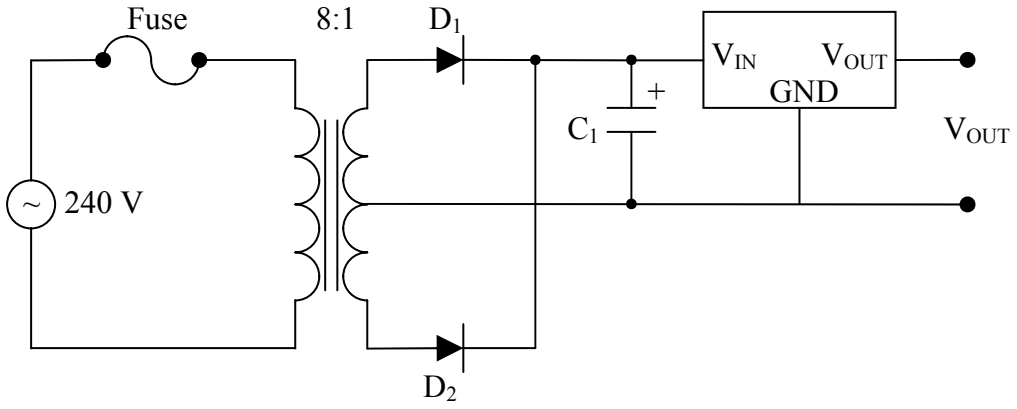
You should recognise the various parts of the circuit and understand their function. The power supply takes a 240 V A.C. mains input and applies this to the primary winding of a 20:1 transformer through a fuse. The fuse is there to protect the circuit if too much current is drawn, either by overloading the output or due to a circuit fault. The transformer steps the voltage down to 12 V RMS. This voltage is rectified by a full-wave bridge rectifier consisting of diodes D_1 – D_4. The resulting waveform is passed through a low-pass filter consisting of inductor L_1 and smoothing capacitor C_1, which reduce the ripple to an acceptably low level.

The inductor also provides “inrush protection”, preventing the transformer from being damaged by the high current that would flow when the supply was first turned on, when the capacitor is completely discharged, if there was no inductor. Instead of allowing a very high current to flow initially, the inductor opposes the change in current, allowing it to build up more gradually. In practise this inductor is often omitted in simple designs, with the self-inductance of the transformer secondary being sufficient to prevent damage.

A Regulated Power Supply

Although the simple power supply is quite practical, it has two weaknesses. First, although the ripple is significantly attenuated by the low-pass filter on the output, some ripple will remain, which may cause problems with sensitive equipment. Second, although the output voltage is nominally 12 V, in practise it will vary between 11 and about 16 V depending on the circuit load, which again may cause problems for sensitive equipment.

Both of these problems can be solved by adding a voltage regulator to the basic design. Although a Zener diode could be used as a voltage regulator, they are only suitable for low-current applications, so they are typically used to stabilize a reference voltage that is then used to control the output voltage of the voltage regulator. Although the entire regulator can be (and often is) made from discrete components like diodes, capacitors and transistors, we will use an integrated circuit that is specifically designed as a voltage regulator. An integrated circuit consists of a number of different electronic devices all fabricated (made) and interconnected together on a single wafer of silicon. They are available to perform many common tasks, including voltage regulation.



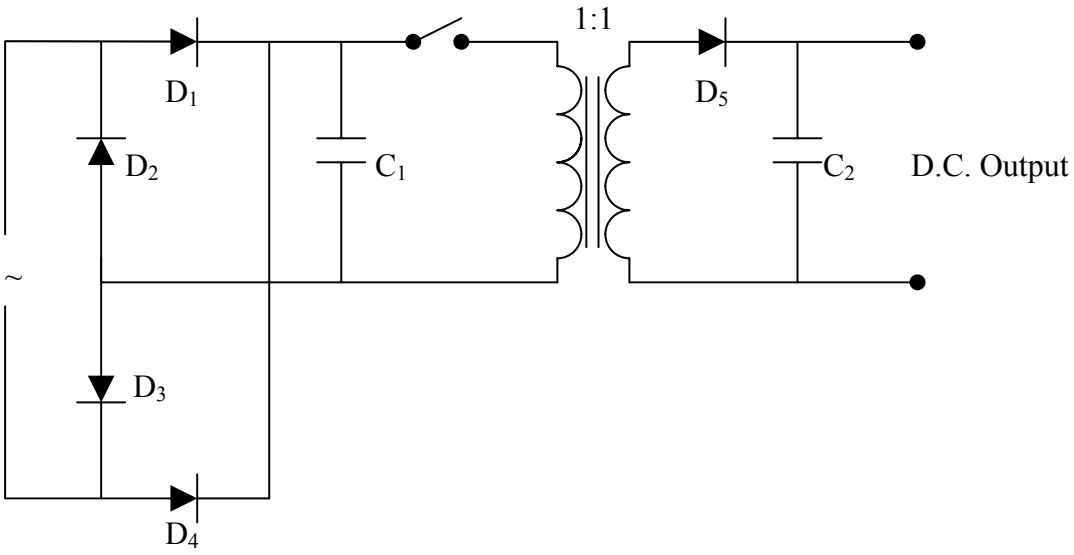
You will notice that the rectifier design is different. Instead of using four diodes in a bridge circuit, this design only uses two diodes. However it still achieves full-wave rectification by making use of a *centre tap* on the secondary winding of the transformer. This is just a separate connection to the middle of the secondary winding. This allows the secondary to function almost like two separate windings. On each half cycle either D_1 or D_2 will conduct, but not both. Whichever diode conducts will connect the positive side of the transformer to the positive side of C1. Current flowing back to the centre tap of the secondary completes the circuit. In effect, only half of the secondary winding is used in each half cycle.

The voltage regulator has three terminals labelled V_{IN}, V_{OUT} and GND (for "ground"). It acts a bit like a variable resistor that is connected between the V_{IN} and V_{OUT} terminals and which is continually adjusted to maintain the voltage between the V_{OUT} and GND terminals at a constant level, say 12 V. As well as regulating the voltage, the voltage regulator also substantially reduces the ripple, since it is able to change its internal resistance fast enough to counteract the voltage fluctuations due to the ripple, thus maintaining a good "clean" D.C. supply despite considerable ripple in the input voltage. For this reason We have also simplified the low-pass filter by removing the inductor, leaving only the smoothing capacitor C_1, which should provide adequate ripple rejection when used with a voltage regulator integrated circuit.

Most integrated voltage regulators provide another benefit: they automatically limit the current and power dissipation of the regulator to safe limits, avoiding damage to the power supply even if the load is short circuited. Despite this the mains supply of a power supply (or any other mains powered device) should always be fused in case of a short circuit within the power supply itself.

Switching Power Supplies

Power supplies that use a transformer to reduce the voltage followed by a rectifier and a voltage regulator are called *linear* power supplies. An alternative design is the *switching* power supply. Instead of using a transformer to reduce the voltage, these supplies rectify the mains voltage to generate a high voltage D.C. voltage source. This is then switched on and off at a high frequency using a fast solid-state switch, and the resulting waveform is fed through a low-pass filter to filter out the A.C. switching components.



Simplified Circuit Diagram of a Switching Power Supply

The 240V A.C. mains supply is rectified by the full-wave bridge rectifier consisting of diodes D_1 to D_4 and smoothed by C1 to generate 338 V D.C. This is switched on and off at a frequency of 100 kHz or so by a high-speed electronic switch, which is shown in the circuit diagram as a switch. The resulting high frequency A.C. waveform is fed into the primary of the isolating 1:1 transformer. The purpose of this transformer is to prevent the D.C. output from being connected to the mains supply, rather than to perform any voltage conversion. The voltage on the secondary of the isolating transformer is half-wave rectified by D5 and smoothed by C2 to give a D.C. output.

The output voltage of the power supply depends on how much time the switch spends in the "on" position compared to how much time it spends in the "off" position. If the switch spends only a small percentage of its time in the "on" position, then only a little power will be transferred to the transformer, and the output voltage will be small. If it spends a lot of time in the "on" position then a lot of power will be transferred and the output voltage will be high. In actual practice, the amount of time that the switch spends in the "on" position is controlled by electronics (not shown n the diagram) that continually monitors the output voltage and adjusts the duty cycle of the switch in order to maintain a constant output at the desired voltage; this is an example of *feedback*.

All this is pretty complex compared to a simple linear supply, so there must be some significant advantages to make it worthwhile. The main advantages of the switching supply are that

1. It dissipates very little power. The switch is either "on", in which case there is current flowing through it but little voltage across it; or "off" in which case there is a high voltage across it but no current flowing through it. In either case the power dissipation is minimal compared to the linear voltage regulator, which has a voltage drop across it and a current flowing through it at the same time, and so is continuously dissipating power.

2. Because the transformer in a switching supply operates at a very high frequency - usually around 100 kHz instead of the 50 or 60 Hz of a standard mains supply - it can be physically very small and light. This is because the size of the core required in a transformer decreases as the frequency increases.

As a result, switching power supplies are generally smaller, lighter and more efficient than their linear counterparts, and can often run off any mains voltage without having to change a selector switch. However they also have their disadvantages. In particularly, poorly designed switching supplies can generate a lot of radio frequency interference, which is a real problem for amateur radio applications. However well designed and properly shielded switching supplies do not necessarily cause interference. Because of their high power requirements and space limitations, virtually all personal computers use switching power supplies (and in fact some of these supplies can be adapted as general purpose power supplies for amateur use).

Please note that switching supplies are very difficult to design and build and can be quite dangerous due to the high voltages they use, and the switching circuitry is often connected directly to the mains input supply. Although the linear power supplies in earlier sections make good projects for amateurs, the design and construction of switching power supplies should be left to professionals!

Summary

Linear power supplies use transformers to reduce the mains voltage, rectifiers to convert the A.C. to D.C. and an output filter including a smoothing capacitor to reduce the ripple to acceptable levels. In power supplies that use a half-wave rectifier the ripple is at the same frequency as the mains, while in power supplies that use full-wave rectifiers the ripple is at twice the mains frequency.

A voltage regulator serves two purposes: to keep the output voltage constant despite fluctuations in the input voltage or load current; and to further reduce ripple. Integrated circuit voltage regulators may also limit current and power dissipation by the regulator to safe levels.

Switching power supplies work by rectifying the mains supply and then switching it on and off at a high frequency. The voltage output is regulated by changing the duty cycle of the switching waveform; that is, the percentage of the time that the switch is “on”. Although most switching supplies still use transformers to isolate the output from the mains supply, these transformers can be small and light because they operate at high frequency, typically around 100 kHz. The advantages of switching supplies are that they are smaller, lighter and more efficient than linear supplies. However poorly designed switching supplies can generate a significant amount of radio frequency interference.

Revision Questions

1 The ripple frequency appearing at the output of an AC-fed power supply using a full wave rectifier will be:

a. Twice the input frequency.
b. Half the input frequency.
c. The same as the input frequency.
d. Dependent on the number of rectifier diodes.

2 To obtain a full-wave rectified output from a transformer using two diodes the transformer must be:

a. An isolation transformer.
b. A step-down transformer.
c. Centre tapped on the secondary winding.
d. Earthed.

3 By introducing a smoothing capacitor and an inductor in a power supply output:
a. The output voltage will increase.
b. The load can be increased.
c. The output voltage will be regulated.
d. The ripple will be reduced.

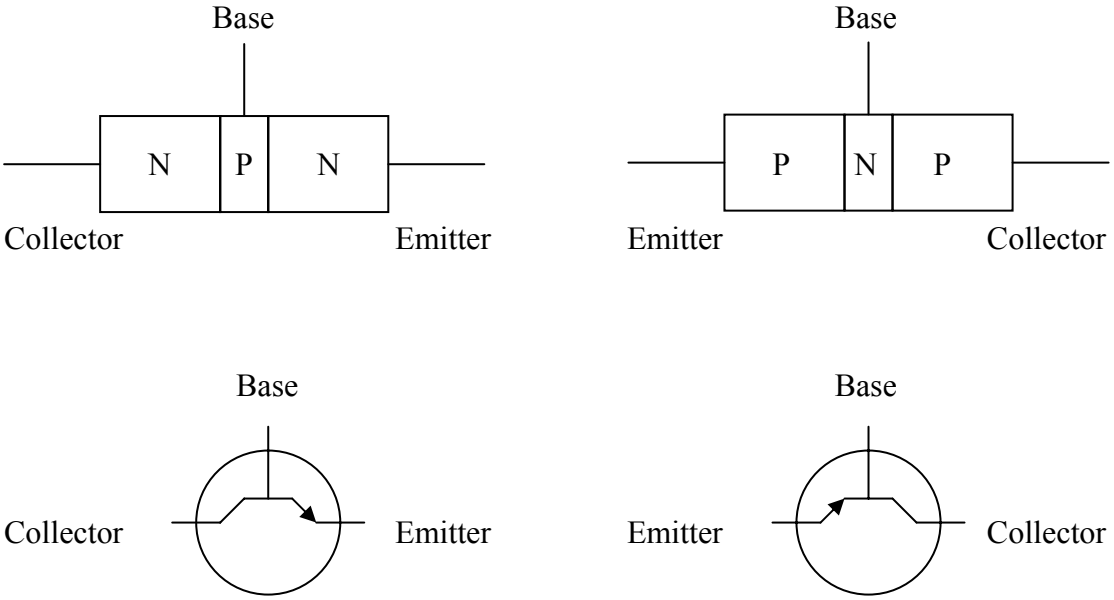
4 A smoothing circuit using an inductor and capacitor is a standard:
a. Low pass filter.
b. Voltage regulator.
c. Rectifier.
d. Discriminator.

5 A voltage regulator in a power supply:
a. Introduces a continuous ripple signal.
b. Allows large currents to be supplied.
c. Protects connected loads from short circuits.
d. Stabilizes the output voltage of the power supply.

6 A zener diode is used in a power supply to:
a. Stabilize a reference voltage.
b. Load an output circuit.
c. Introduce a noise signal.
d. Prevent excessive current flow.

Chapter 16 - The Bipolar Junction Transistor

A bipolar junction transistor ("transistor" for short) consists of a thin layer of P-type or N-type semiconductors called the "base", which is sandwiched between two thicker layers of semiconductor, the "collector" and "emitter" with the opposite polarity. Transistors come in two polarities: NPN transistors, which have a P-type base sandwiched between an N-type emitter and collector; and PNP transistors, which have an N-type base sandwiched between a P-type emitter and collector. The construction and circuit symbols of NPN and PNP transistors are shown below:



The Physical Structure and Circuit Symbol for NPN and PNP Transistors

Note that the terminal with the arrowhead in the circuit symbol is always the emitter. The arrow represents the base/emitter junction, which has similar properties to a diode, and shows which direction the Base/Emitter and Collector/Emitter current flows in.

Operation of the NPN Transistor

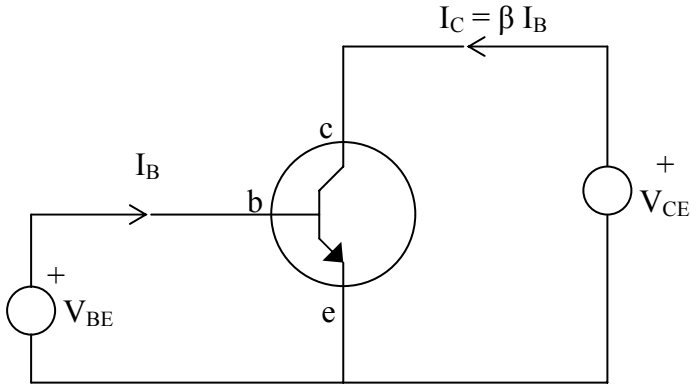
To understand the operation, assume that an NPN transistor like the one on the left has a potential difference applied between the collector and emitter, making the collector positive with respect to the emitter. Then at the junction between the collector and the base, an N-type semiconductor meets a P-type semiconductor, creating a reverse-biased diode junction. Just as in a diode, free electrons from the N-type material will migrate across the junction, filling the holes in the P-type material, and creating a thin depletion layer that will prevent current from flowing.

Now suppose a potential difference is applied between the base and the emitter, making the base more positive than the emitter. The P-N junction acts like a forward-biased diode, so provided the base-emitter voltage exceeds to forward bias voltage for this junction (about 0,6 V for silicon transistors, and 0,2 V for germanium transistors) a current can flow from the base to the emitter. This current is carried by electrons from the emitter that are attracted by the positive potential of the base and cross the junction into the base where they combine with holes.

However because the base is very thin (much thinner than shown), many of the electrons from the emitter that are attracted by the base potential do not end up colliding with holes and recombining, but instead make it all the way across the base and into the collector. Since electrons are now moving from the emitter to the collector, there is a current flow from the collector to the emitter despite the reverse-biased collector/base junction! This is possible because electrons from the emitter that escape recombining with holes in the base act as charge carriers in the depletion layer of the reverse-biased junction, allowing a current to flow.

So in an NPN transistor, making the base 0,6 V or so more positive than the emitter will allow a current to flow both from the base to the emitter and from the collector to the emitter, provided of course that the collector is also positive with respect to the emitter. It turns out that transistors can be designed so that the current that flows from the collector to the emitter (known as the "collector" current) is many times greater than the current from the base to the emitter (the "base" current). This enables transistors to make small signals larger, a process called amplification.

The ratio of collector current to base current is known as the beta or "current gain" of the transistor, and is represented by the Greek letter beta (β). In ordinary small signal transistors (designed for low-power work), β typically ranges from 100 to 500. One limitation is that β is not well controlled, so two transistors from the same batch may have quite different values of β. For this reason it is important to design circuits that do not rely on a particular value of β for correct operation.



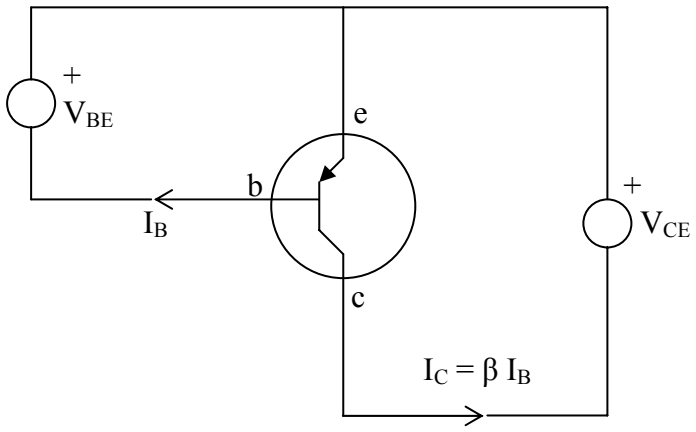
Operation of the NPN Transistor

The operation of the NPN transistor can be summarized as follows:

- The collector should always be kept positive with respect to the emitter.
- If the base/emitter voltage V_{BE} is less than 0,6 V then no base or collector current will flow, and the transistor is "shut off".
- Once V_{BE} reaches 0,6 V, a base current I_B will flow, causing a larger collector current $I_C = \beta I_B$ to flow.
- V_{BE} will remain around 0,6 V as long as any base current is flowing.
- The value of β ranges between 100 and 500 for typical small-signal transistors, but is not well controlled and should not be assumed to have a particular value.

Operation of the PNP Transistor

The PNP transistor operates similarly to the NPN transistor, but with the opposite polarity. When the base gets 0,6 V more *negative* than the emitter a base current I_B and a larger collector current $I_C = \beta I_B$ will flow.



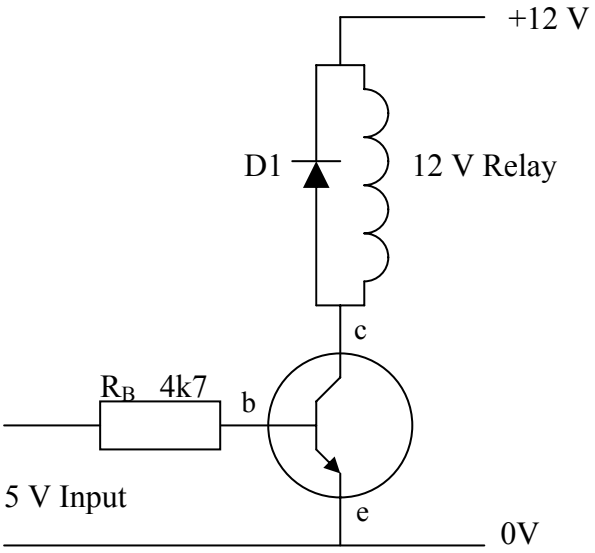
Operation of the PNP Transistor

The operation of the PNP transistor can be summarized as follows:

- The collector should always be kept negative with respect to the emitter.
- If the base/emitter voltage V_{BE} is greater than -0,6 V then no base or collector current will flow, and the transistor is "shut off".
- Once V_{BE} reaches -0,6 V, a base current I_B will flow, causing a larger collector current $I_C = \beta I_B$ to flow.
- V_{BE} will remain around -0,6 V as long as any base current is flowing.
- The value of β ranges between 100 and 500 for typical small-signal transistors, but is not well controlled and should not be assumed to have a particular value.

The Transistor Switch

Transistors are often used as switches because a small base current can turn the transistor "on" and allow a large collector current to flow. For example, suppose you want to switch a 12 V relay that draws 20 mA from a 5 V microprocessor control signal that can only supply 1 mA. You could use the following circuit:



When the input signal is off (0 V), no base current flows so the transistor is turned off and no collector current flows. When the input signal is turned on (+5 V) it raises V_{BE} to 0,6 V so the voltage across the base resistor R_B is 5 – 0,6 = 4,4 V, so the current flowing through R_B and into the base I_B = 4,4 / 4700 = 0,94 mA. This is sufficient to turn the transistor "on", causing a collector current to flow and operate the 12 V relay. The Diode D1 prevents the back EMF from the inductance of the relay coil from destroying the transistor when the relay is turned off again.

Transistor switches like this are usually used to switch D.C. voltages and currents. When high frequency signals need to be switched, diodes or relays are usually used. Of course the relay driver circuit might be similar to the one above.

Summary

Bipolar Junction Transistors are semiconductor devices that consist of a thin *base* made of either P- or N-type material sandwiched between a *collector* and *emitter* made of the opposite polarity semiconductor.

In an *NPN* transistor, both the collector and base are made positive with respect to the emitter. If the base/emitter voltage is less than 0,6 V the transistor is turned off and no collector current will flow. When the base/emitter voltage reaches about 0,6 V, a small base current and a large collector current will flow.

In a *PNP* transistor, both the collector and base are made negative with respect to the emitter. If the base/emitter voltage is greater than -0,6 V the transistor is turned off and no collector current will flow. When the base/emitter voltage reaches about -0,6 V, a small base current and a large collector current will flow.

The ratio of the collector current to the base current is called the "beta" of the transistor, and typical values for small-signal transistors range between 100 and 500. As long as any base current is flowing, the base/emitter voltage will remain around 0,6 V for NPN transistors, or –0,6 for PNP transistors, irrespective of the actual base or collector current. Since transistor betas may vary widely, even for transistors of the same type, circuits should not rely on a specific value of beta.

Transistors can be used as D.C. switches, to allow a low voltage or small current (or both) to switch a larger voltage and current.

Revision Questions

1 If the base potential of a NPN transistor is held at the emitter potential, the collector current will be:

a. Zero.
b. Always 1 A.
c. Between 10 mA and 2 A.
d. Very high.

2 For a silicon transistor to conduct:

a. The base-emitter must be forward biased by 0,6 V.
b. The base must be connected to the emitter.
c. The collector must be connected to the emitter.
d. The base lead must be disconnected.

3 The beta of a transistor is:

a. The ratio of the collector current to the base current.
b. The ratio of the collector voltage to the base voltage.
c. The ratio of the collector current to the emitter current.
d. The ratio of the collector voltage to the emitter voltage.

4 For a collector current to flow in a PNP transistor:

a. Both collector and base must be positive with respect to the emitter.
b. Both collector and base must be negative with respect to the emitter.
c. The collector must be positive and the base negative with respect to the emitter.
d. The collector must be negative and the base positive with respect to the emitter.

5 If a transistor is used to control a relay that does not have protection diodes, then the back EMF from the relay solenoid may:

a. Increase the switching time.
b. Decrease the switching time.
c. Deduce the power consumption.
d. Damage the transistor.

6 If the base current in a transistor is 100 µA and the beta of the transistor is 100, then the collector current is:

a. 1 mA.
b. 10 mA.
c. 100 mA.
d. 1 A.

7 For an NPN transistor in normal operation:

a. The collector voltage exceeds the emitter voltage.
b. The emitter voltage exceeds the collector voltage.
c. The emitter voltage exceeds the base voltage.
e. The collector and emitter voltages are equal.

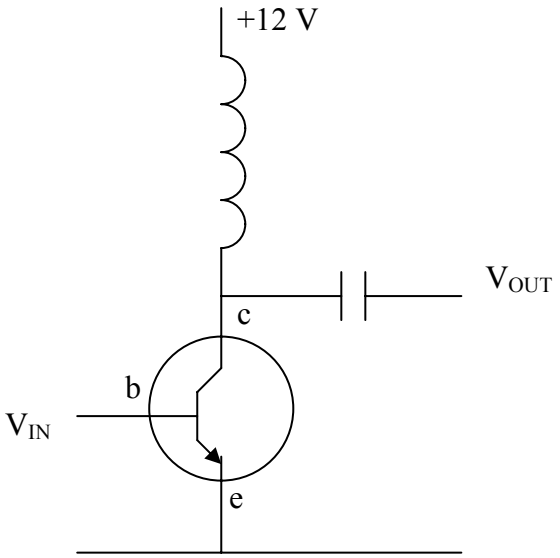
Chapter 17 - The Transistor Amplifier

Amplification is the process of increasing the power of a signal. Since power is V * I, this will involve increasing either the voltage, or the current, of the signal, or possibly (but not necessarily) both. It is quite possible to amplify a signal without increasing the voltage of the signal, provided that the current is increased. Conversely, it is possible to increase the voltage of a signal – perhaps by the use of a step-up transformer – without amplifying it. In the case of the step-up transformer, although the voltage of the signal is increased, the current is decreased so the power remains the same and no amplification takes place.

Amplification is crucial for radio receivers. The radio signals received from the antenna may be as small as –130 dBm (10^{-16} watts). In order to make them audible, the receiver must convert them into signals in the order of 0 dBm (1 mW), which requires amplification by a factor or 10^{13}. In actual fact the amplification of a radio receiver is often greater than this, to make up for losses in other components like filters.

Class C Amplifiers

Because the transistor allows a large collector current to be controlled by a small base current, transistors are used in many amplifiers. Let us consider a simple design for a transistor amplifier to work at radio frequencies.



Here the input signal is applied directly to the base of the transistor and the output signal is taken from the collector. This circuit is called a “common emitter” amplifier, because the emitter of the transistor is common to both the input and output circuits.

The inductor and capacitor play two roles. Firstly, the inductor allows D.C. to pass while blocking radio frequency (RF) signals. This allows the collector of the transistor to be *biased* so that it is positive with respect to the emitter, as required for correct operation. It also provides the route for supply current to flow from the +12 V source to the amplifier – since we are making the signal more powerful, and power cannot be manufactured from thin air, it must come from somewhere. In this case it comes from the +12 V supply connected via the inductor to the collector. The output capacitor allows the amplified A.C. signal to pass, while blocking the D.C. supply voltage, preventing the bias voltage from interfering with the stage that follows this amplifier.

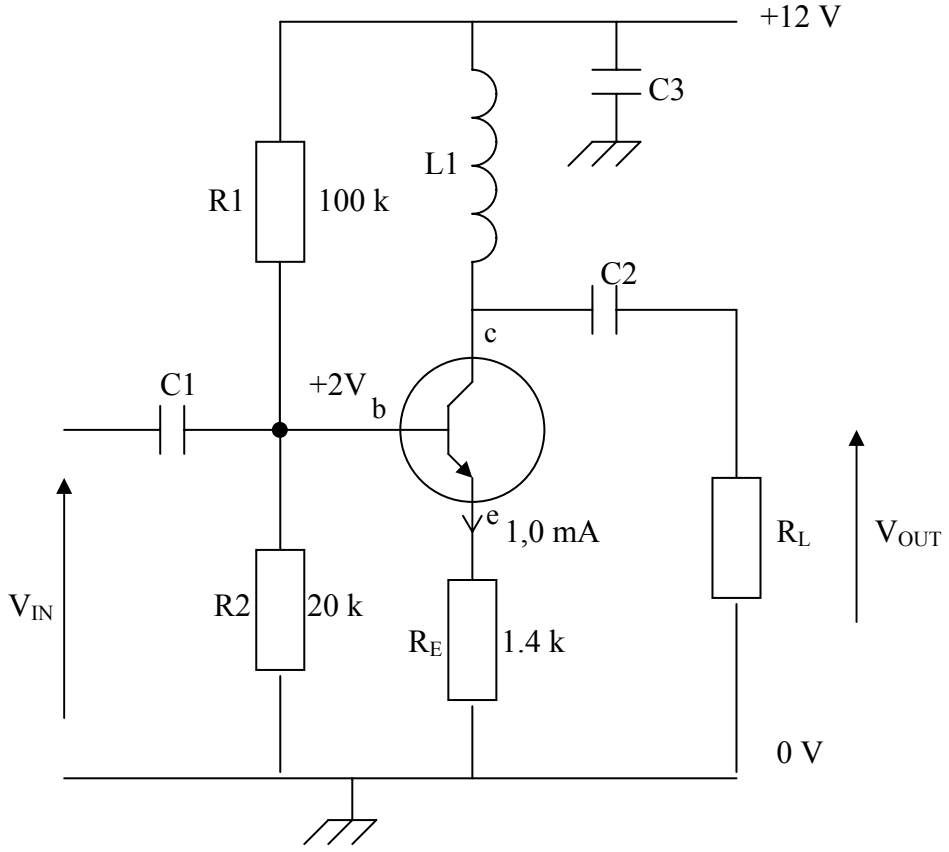
Let us consider the operation of this circuit. If the input signal is less than 0,6 V, the base-emitter voltage for a silicon transistor, then no current will flow into the collector, and the collector voltage will be +12 V. If the input signal exceeds 0,6 V, then the transistor will start to conduct, with the collector current equal to the transistor beta times the base current. The inductor will oppose any sudden change in the flow of current by generating a voltage across itself that will reduce the collector voltage. When the input signal drops below 0,6 V, the transistor will stop conducting. By this time there will be some current flowing through the inductor, and the inductor will oppose any sudden change to the flow of current by generating a voltage across itself in the other direction, this time raising the collector voltage above the 12 V supply voltage, possibly to as much as 24 V (twice the supply voltage). When the input signal again exceeds 0,6 V and the transistor starts to conduct, the collector voltage will again drop, and so on. These variations in the collector voltage constitute an A.C. signal that will be passed through the capacitor as an output signal.

So in this amplifier, the transistor does not conduct the whole time. In fact, it conducts somewhat less than half of each A.C. cycle, since it will only start to conduct when the input voltage exceeds 0,6 V. An amplifier that conducts for less than one half of a cycle is called a "Class C" amplifier. Since class C amplifiers do not reproduce the shape of the input waveform accurately, they generate a large amount of distortion. However they do have the advantage of being quite efficient compared to other amplifiers, since most of the power supplied to the amplifier (in this case from the +12 V supply) ends up in the output signal. Efficiencies of 60 to 70% are common for Class C amplifiers. Also note that there is not much point trying to amplify a signal with a peak voltage of less than 0,6 V using this amplifier, since it won't ever be sufficient to start the transistor conducting!

So what would one do with an amplifier that requires a large input signal and generates considerable distortion, but is relatively efficient? Class C amplifiers are often used as the RF power amplifier for CW (Morse code) and FM transmitters, since in these applications it turns out that the distortion (also called *nonlinearity*) of the amplifier is not a big problem as the unwanted harmonics can be filtered off quite easily by a low-pass filter at the output. However Class C amplifiers are *not* suitable for use in single sideband (SSB) or amplitude modulation (AM) transmitters; for these you need a *linear* amplifier. (Don't worry if you are not familiar with the terms CW, AM, FM and SSB, they will be explained in a later module).

The Class A Common-Emitter Amplifier

In order to faithfully amplify small signals, we usually use Class A amplifiers. In this class of amplifier, collector current flows in the transistor throughout the entire cycle of the input waveform. A typical circuit might be as follows:



Here R1 and R2 form a *voltage divider* that applies a certain *bias voltage* to the base. For example, suppose that R1 is 100 k, and R2 is 20 k, then the bias voltage at the base will be 2 V. Since this is greater than 0,6 V it is sufficient to cause a current to flow from the base to the emitter, and we know that the base/emitter voltage will be about 0,6 V. This means the voltage across R_E, the emitter resistor, is 2 V – 0,6 V = 1,4 V. Suppose R_E is 1,4 k, then the current flowing through R_E (and also through the emitter of the transistor) is 1,4 V / 1,4 k = 1,0 mA.

So how much of this is base/emitter current, and how much is collector/emitter current? That depends on the Beta of the transistor. If the beta is 99, then the collector/emitter current is 99 times the base/emitter current, so 1% of the emitter current comes from the base, and 99% from the emitter. In this example, the collector current would be 0,99 mA.

But hold on a moment. We have already been warned that the Beta of a transistor is not well controlled and should not be relied upon to have a specific value. What if the
Beta is 499 instead of 99? Well then the collector current will be 499/500 of the emitter current, or 0,998 mA, instead of 0,990 mA. As you can see, the collector current is constant to within 1%, despite variations in the Beta from 99 to 499. For most practical purposes, the collector current can be assumed to be equal to the emitter current, with the base current being ignored (although of course the base current is very important in practice, as it is what allows the collector current to flow!)

Finally let us consider the input and output impedance of the amplifier. An RF input signal will "see" an input impedance consisting of resistors R1 and R2 in parallel. Why in parallel? Well the supply voltage (in this case +12 V) always has an A.C. path to ground. In this case, it is provided by the "decoupling" capacitor C3, which is connected between the supply voltage and the chassis. So the input impedance is 100 k in parallel with 20 k, or about 16,7 k. (There is an additional component, consisting of the emitter resistor R_E multiplied by the Beta of the transistor that is also in parallel with the input, but this is usually significantly higher than the bias resistors R1 and R2 and can be neglected).

Having determined the input impedance, what then is the output impedance? Well since the collector of the transistor acts like a good current source (the current flowing, Beta times the base current, does not depend much on the collector voltage) the effective collector impedance is fairly high. The inductor L1 will also be chosen to exhibit high reactance at the design frequency. So we can say that in this case the output impedance is "quite high" without actually putting a figure to it.

And the gain of the amplifier? Well assume that we change the base voltage by a small amount, which we shall call V_{IN}. Then since the base/emitter voltage remains constant at 0,6 V, the emitter voltage must also change by the same amount. This will cause a the emitter current I_E also to change by a small amount, $\Delta I_E = V_{IN}/R_E$. But since the collector current is virtually identical to the emitter current (we are neglecting the small effect of the base current), the output current will change by the same amount as the emitter current, so $I_{OUT} = V_{IN}/R_E$. The effect that this has on the output voltage (and hence the gain of the amplifier) will depend on the resistance of the load, R_L. To be precise,

$$\begin{aligned} V_{OUT} &= I_{OUT} R_L \\ &= (V_{IN}/R_E) R_L \\ &= V_{IN} R_L/R_E \end{aligned}$$

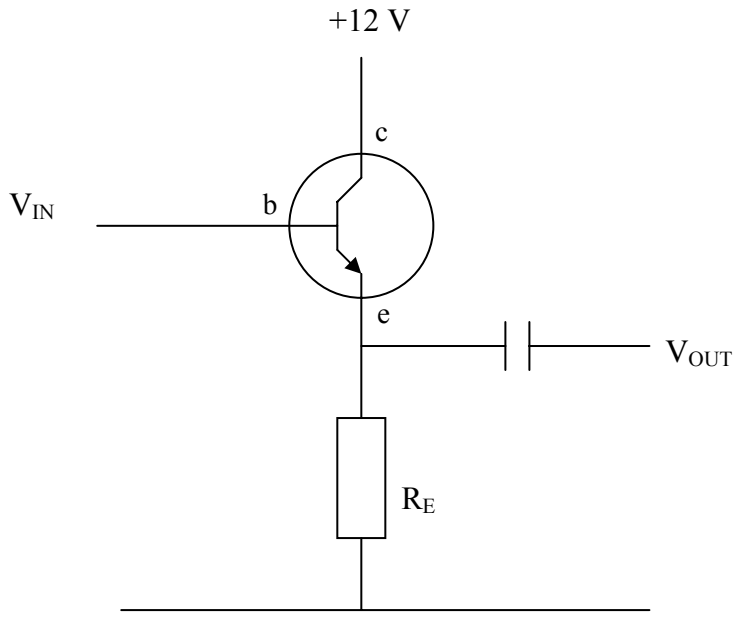
So the *voltage gain* of this amplifier is R_L/R_E. To calculate the *power gain*, we need a specific value for the load resistance. Let's say it is 16,7 k, the same value as the input resistance of the amplifier. Then the power gain is just the square of the voltage gain. The voltage gain will be 16,7 k / 1,4 k = 11,9 and the power gain will be the square of this, 142 or 21,5 dB. This is typical of the gain achievable by a small-signal common-emitter amplifier.

Note that in this design, because the base of the transistor is biased to a voltage of +2 V, the transistor will conduct provided the input voltage does not go below -1,4 V. So as long as the input voltage is less than 1,4 V *peak* (about 1,0 V RMS), the transistor will conduct throughout the full cycle of the input waveform, making this a Class A amplifier. It will faithfully reproduce the exact shape of the input signal at the output, and so is a *linear* amplifier that can be used to amplify SSB and AM signals without distortion.

Of course, the DC bias current that flows through the collector circuit of the transistor the whole time does waste quite a lot of power, making Class A amplifiers much less efficient than their Class C counterparts. Class A amplifiers are typically only about 25% efficient – that is, only 25% of the power supplied by the power supply actually ends up in the load. The other 75% ends up heating the output transistor!

The Common-Collector (Emitter Follower) Amplifier

Consider the following circuit:



Suppose that the input voltage is always between 0,6 V and 12 V, so the transistor always conducts (in other words, it is operating Class A). This time, the output voltage is taken from the emitter instead of the collector of the transistor. What do we know about the output voltage? Well since the base/emitter voltage is always 0,6 V if the transistor is conducting, the emitter voltage must "follow" the base voltage, although it will always be 0,6 V less than the base voltage. So any change in the input voltage will result in an equal size change in the output voltage, and the amplifier has a voltage gain of 1. (Note that the DC bias is removed by the DC blocking capacitor at the output, so the output voltage is *not* necessarily 0,6 V below the input voltage.)

This circuit is called a "common collector" amplifier, and is also known as an "emitter follower" because the emitter voltage "follows" the base voltage.

So why would anyone want an amplifier if the output voltage is the same as the input voltage? The secret is in the impedances. The output impedance of the emitter follower is low, making it a good voltage source (because the output voltage ill not depend on the load resistance). However the input impedance is high, so it does not load the stage preceding it much. This makes it a good circuit to use as a buffer stage, to prevent changes to the input impedance of the stages following it from affecting the stages that precede it. And since the low-impedance output is capable of supplying much more power than the high-impedance input consumes, the common collector amplifier is quite capable of providing a *power* gain even though it has no *voltage* gain.

The Common Base Amplifier

There is a third transistor amplifier configuration, known as the "common base" configuration, which is less common than the common emitter or common collector configurations. It has low input impedance (typically only a few ohms) and high output impedance (making is a good current source), and a current gain of unity (one). You could

consider it to be a "current follower" since the output current is identical to the input current. It can also provide a power gain if the load resistance is greater than the input impedance of the amplifier.

The Class AB Amplifier

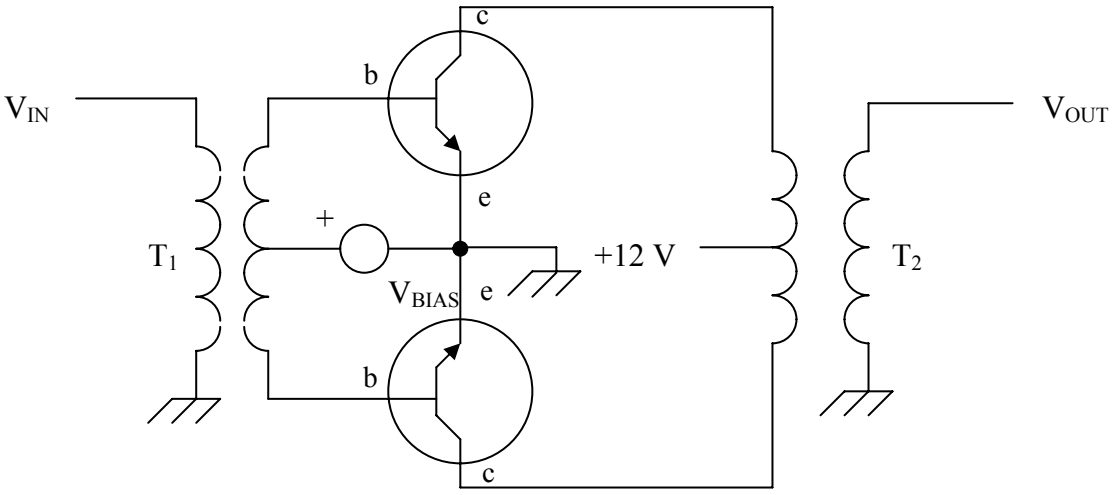
In a Class A amplifier, the transistor (or other output device) conducts for the full cycle of the input waveform. This is also referred to as conduction over 360˚ of the input cycle. Class A amplifiers are linear – that is, they accurately reproduce the shape of the input waveform at the output and so introduce little distortion – but inefficient.

In a Class C amplifier, the transistor conducts for less than half the cycle, in other words for less than 180˚. Class C amplifiers can be pretty efficient, but they are very non-linear, introducing a substantial amount of distortion into the output.

There is also a Class "B", where the transistor conducts for exactly half the input form (180˚). However this is not commonly used.

Class AB amplifiers use two output devices operating in a "push-pull" configuration were one device conducts during the positive half cycle of the input waveform, and the other device conducts during the negative half cycle. Both devices are operated Class B, meaning that they conduct for half of the input waveform (180˚). However, between them the output devices can replicate both the positive and the negative half cycles of the input waveform, as can a Class A amplifier, which is where the name "Class AB" comes from – two Class B devices operating together to provide the same effect as a single Class A device.

A simplified push-pull Class AB amplifier circuit is shown below:



In this circuit, the bias voltage V_{BIAS} is just sufficient to keep both transistors just conducting a small current when there is no input voltage. The transformer at the input acts as a "phase splitter" to convert the input signal into two signals that are 180˚ out of phase, which are applied to the bases of the transistors. When a positive signal is applied to the base of one transistor is conducts more strongly, while the voltage at the base of the other transistor is reduced so it stops conducting. The transistors switch around every half cycle, so each of the transistors is conducting for half the cycle, or about 180˚. The transformer at the output recombines the output of the two transistors, so although each is only conducting for half the cycle, the combined output of both transistors can faithfully replicate both half cycles of the input signal.

Class AB amplifiers have the advantage that because there isn't a large constant DC bias current flowing through the output devices (as there is in Class A) they are much more efficient than Class A amplifiers; while because they reproduce both the positive and the negative half cycles of the input, they are much more linear than Class C amplifiers. Although they still introduce some distortion, such as *crossover* distortion at the point where one of the transistors stops conducting and the other starts conducting, this can be kept within reasonable limits, so properly designed Class AB amplifiers are quite suitable for use as power amplifiers for AM and SSB signals.

Summary

An amplifier is a circuit that increases the power of a signal.

The Common Emitter amplifier can have both voltage and current gain. Common emitter amplifiers have high output impedance and moderate (10 k or so) input impedance.

The common collector amplifier is also known as the "emitter follower" because the output is taken from the emitter, which "follows" the voltage on the base. The common collector amplifier has unity voltage gain but can still provide power gain if the output current is greater than the input current. The common collector (emitter follower) has a high input impedance and low output impedance.

The common base amplifier has a low input impedance and high output impedance, and has a current gain of unity (i.e. the output current is the same as the input current).
Class A amplifiers conduct for the full cycle (360˚). They have low distortion (good linearity) but are relatively inefficient. Almost all small-signal amplifiers are Class A, because efficiency is not important.

Class B amplifiers conduct for exactly half the cycle. They are not commonly used.

Class C amplifiers conduct for less than half the cycle (less than 180˚). They can be very efficient, but are non-linear and introduce distortion. Although they can be used as power amplifiers for CW and FM signals, they cannot be used for AM or SSB signals.

Class AB amplifiers use two Class-B output devices operating in push-pull configuration to amplify both the positive and the negative half cycles. They are more efficient than Class A amplifiers, and while they also have more distortion than Class A amplifiers they can still be used as power amplifiers for AM and SSB signals.

Revision Questions

1 The output impedance of an emitter follower buffer amplifier is:
a. Infinite.
b. Very high.
c. 0.
d. Fairly low.

2 In a transistor amplifier circuit where full base current always flows, the circuit is biased for:
a. Class A amplifier.
b. Class B amplifier.
c. Class AB amplifier.
d. Class C amplifier.

3 A class C amplifier conducts over:
a. The complete cycle.
b. Three quarters of the cycle.
c. Exactly half a cycle.
d. Less than half a cycle.

4 The amplifier class which has the lowest distortion figures is:
a. Class A.
b. Class B.
c. Class AB.
d. Class C.

5 An amplifier that operates under conditions of bias and supply such that conduction occurs for more than 180 degrees but less than 360 degrees of a complete input cycle is operating in class:
a. Class A.
b. Class AB.
c. Class B.
d. Class C.

6 When an RF power amplifier is biased for a conduction angle of more than 360 degrees:
a. Output current flows for only part of the input cycle.
b. Bias current never shuts off the device.
c. The average grid voltage is twice cutoff voltage.
d. RF power is produced at greatest efficiency.

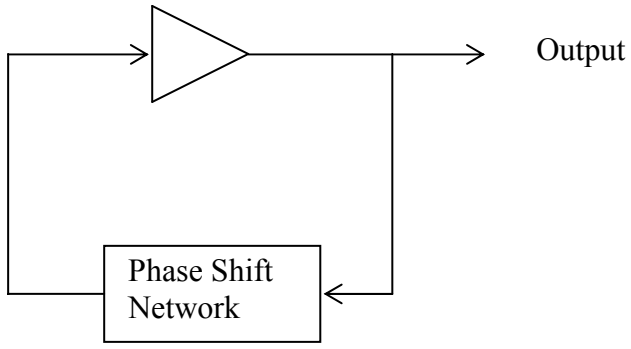
Chapter 18 - The Oscillator

Oscillators are circuits that are used to generate A.C. signals. Although mechanical methods, like alternators, can be used to generate low frequency A.C. signals, such as the 50 Hz mains, electronic circuits are the most practical way of generating signals at radio frequencies.

Oscillators are widely used in both transmitters and receivers. In transmitters they are used to generate the radio frequency signal that will ultimately be applied to the antenna, causing it to transmit. In receivers, oscillators are widely used in conjunction with mixers (a circuit that will be covered in a later module) to change the frequency of the received radio signal.

Principal of Operation

The diagram below is a *block diagram* showing a typical oscillator. Block diagrams differ from the circuit diagrams that we have used so far in that they do not show every component in the circuit individually. Instead they show complete functional blocks – for example, amplifiers and filters – as just one symbol in the diagram. They are useful because they allow us to get a high level overview of how a circuit or system functions without having to show every individual component.



Block Diagram of an Oscillator

The triangular symbol at the top represents an amplifier. The input of the amplifier is the blunt side of the triangle, on the left in this diagram; the output is the pointy side of the triangle. Since this symbol always represents an amplifier, there is no need to label it. The output of the amplifier is connected to the input of the block labeled "phase shift network", and the output of the phase shift network is connected back to the input of the amplifier. (Since the rectangular box of the phase shift network does not indicate the input and output, you must surmise this from the directions of the arrows on the connecting lines.) The output of the oscillator is taken from some point in the circuit - in this diagram We have shown it being taken from the output of the amplifier.

The lines connecting the symbols in the block diagram represent the flow of signals from one functional block to another. In this type of diagram, a line does not necessarily represent a single wire, as it would in a circuit diagram. A signal might flow along a single wire (with respect to earth), or it might flow in two wires, with the current flowing in opposite directions in both wires. In either case, it could be represented by a single line in a block diagram. The arrows at the end of the lines show the direction that the signal flows in – in this case, from the output of the amplifier to the input of the phase shift network, and from the output of the phase shift network back to the input of the amplifier. The direction in which the signal is flowing does not in general correspond with the direction in which *current* is flowing – after all, most of the signals we deal with will in any case be A.C. so current flows in *both* directions.

So how does this circuit oscillate? When it is initially turned on, there will be some (very small) *thermal noise* present in the circuit. This type of noise is generated by the random motion of electrons due to heat, and exists in all conductors. Thermal noise is broadband in nature, meaning that it includes frequency components at all possible frequencies. (When you turn the volume of a hi-fi amp up without any input signal, the hiss you hear is the audio frequency component of the thermal noise. If you hear a hum, this is mains pickup, not thermal noise.)

Thermal noise at the input to the amplifier will be amplified, causing a larger noise signal at the output of the amplifier, some of which is bled off to the output, and some of which is applied to the phase shift network. The phase shift network does what its name implies – it changes the phase of the input signal, so the output of the network will have a phase that either leads or lags the input signal. The phase relationship between the output and the input depends on the precise frequency of the input signal.

At most frequencies, the output of the phase shift network, which is fed back into the amplifier, will not be at precisely the same phase as the noise component that caused it in the first place. In this case, the signal that is "fed back" to the input of the amplifier from the phase shift network will partially cancel out the signal that caused it, so the noise components at these frequencies will die out. However at one frequency, the output of the phase shift network will be exactly in phase with the noise component that caused it, and so it will reinforce that particular frequency component of the noise signal at the input to the amplifier.

This reinforced signal will again be amplified by the amplifier, phase shifted by the phase shift network, and fed back to the input of the amplifier. And once again, the output from the phase shift network is precisely in phase with the input signal from the "last round" that caused it, and so the signals reinforce each other and keep on growing.

Of course the signal cannot grow larger forever. As the signal grows bigger, ultimately the gain of the amplifier will be reduced (for example, it may be limited by the power supply voltage to the amplifier) until we reach the point that the amplified signal that is passed through the phase shift network and back to the input of the amplifier is only just as strong as the input to the amplifier that caused it. At this point, the signal is no longer growing, but remains constant and we have reached a stable oscillating state. If the oscillator has been designed correctly, then the output will be a constant amplitude signal at the desired frequency.

Feeding back some of the output of the amplifier back to the input in such a way that it reinforces the original input signal is called *positive feedback*. This is the same effect that you get when the audio output of a PA system is fed back to the microphone creating "howl-around" or "feedback".

The Barkhausen Criteria for Oscillation

The *loop gain* of an oscillator is the total gain that the signal experiences starting from any point in the circuit and going around the loop until it gets back to the starting point. For example, suppose the amplifier has a gain of 10 dB, that half the power is "bled off" to the output (resulting in a loss of 3 dB), and that the phase shift network also has a loss of 3 dB. Converting the losses into negative gains, we get the following figures:

Amplifier	10 dB
Loss of output signal	-3 dB
Phase shift network	-3 dB
Total loop gain	4 dB

Similarly, you can calculate the total phase shift around the loop. The amplifier will contribute some phase shift, and the phase shift network will contribute some more. Even the interconnecting wires may contribute significant phase shift at high frequencies – for example, the wavelength of a 100 MHz signal is 3 m, so every centimeter of connecting wire would contribute a phase shift of about 1,2˚!

When the oscillator is oscillating stably – that is, with constant amplitude and frequency, the following criteria must be fulfilled:

- The *loop gain* must be exactly 1. If the gain was more than 1, then the amplitude of the output would be increasing. If less than 1, then the amplitude would be decreasing.

- The *total phase shift* around the loop must be 0 or an integer multiple of 360˚. This is necessary for the signal to reinforce itself as it goes around the loop, so it does not cancel itself out.

These requirements are known as the *Barkhausen criteria* for oscillation.

It is entirely possible for these criteria to be met at more than one frequency. In particular, it is easy for the phase requirement to be met, since it only specifies a phase shift of 0 or any integer multiple of 360˚, so it could be satisfied for different frequencies that had a total phase shift around the loop of say 0˚, 360˚ and 720˚. If both criteria are met for several frequencies, then oscillator will oscillate at all these frequencies simultaneously, which is usually not the desired result! Oscillations at undesired frequencies are called *parasitic* oscillations.

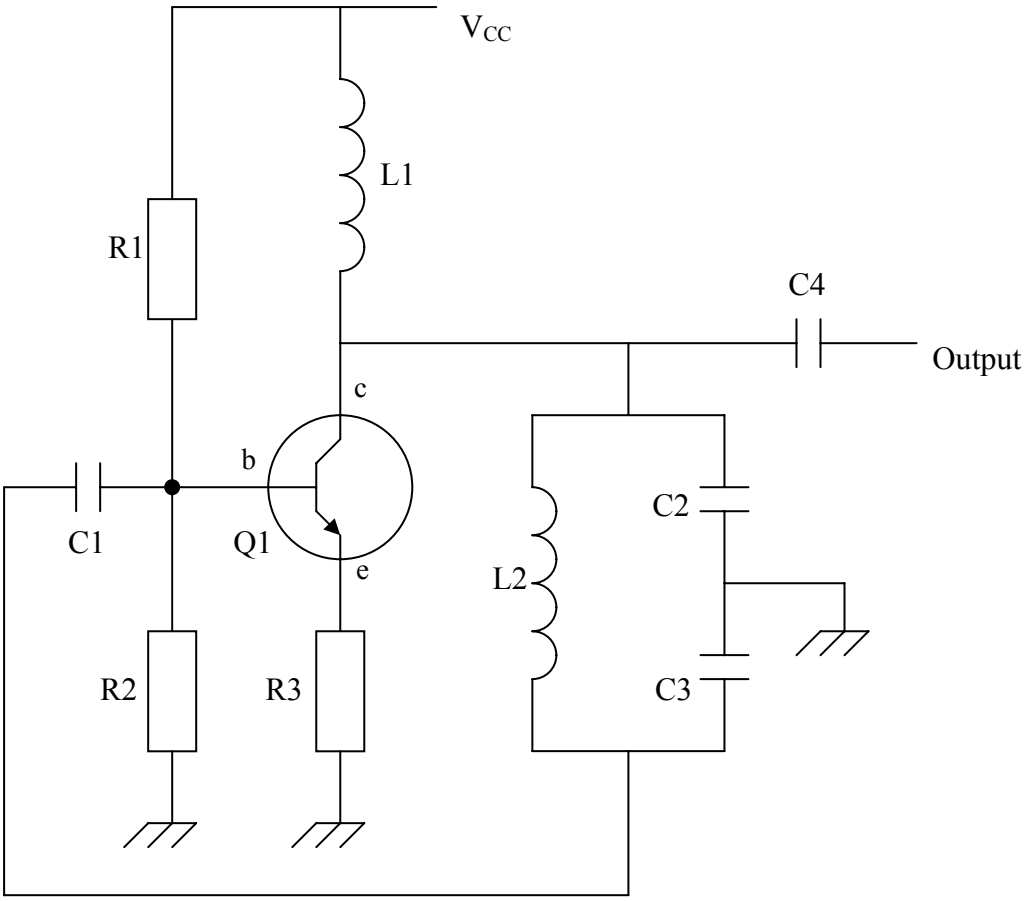
In order to minimize the chance of this happening, the phase shift network is usually also made frequency selective, so that it will pass frequencies in the region of the desired frequency of oscillation, while attenuating frequencies that are higher or lower than this. In other words, it is made to be a *band-pass filter* as well as a phase shift network. The advantage of this is that even if the phase shift criterion is met for some other frequencies, as long as they are far enough away from the desired frequency, they can be attenuated sufficiently by the band-pass characteristic of the network to ensure that the loop gain remains less that 1 so oscillation will not occur at these unwanted frequencies.

Fortunately there is a simple circuit that provides both a phase shift and band-pass filter characteristics simultaneously: the parallel tuned circuit. At the resonant frequency the reactance of a parallel tuned circuit changes rapidly from being highly inductive just below the resonant frequency to being highly capacitive just above the resonant frequency. This sudden change in reactance results in a change in the phase relationship between the voltage across the tuned circuit and the current flowing through it (remember that for in inductive reactance, voltage leads current, while for a capacitive reactance current leads voltage). At the same time, the parallel tuned circuit can be used to provide good band-pass filter characteristics, minimizing the likelihood of parasitic oscillation.

An oscillator that uses a tuned circuit as its phase shift network will oscillate at (or very close to) the resonant frequency of the tuned circuit.

The Colpitts Oscillator

The Colpitts oscillator is typical of how these concepts can be implemented in a practical circuit.



Circuit Diagram of a Colpitts Oscillator

Transistor Q1 and the associated components R1, R2, R3 and L1 form a common-emitter amplifier. The output of the amplifier, taken from the collector of Q1, is fed into a parallel tuned circuit consisting of L2, C2 and C3. The capacitor in this tuned circuit has been "split" into two capacitors, C2 and C3, to allow the output current from the collector of Q1 to flow to ground via C2. This causes a voltage across the whole parallel tuned circuit (also known as the *tank circuit* of the oscillator). This voltage is fed back to the input of the amplifier via C1. The output of the oscillator is taken from the collector of the transistor via C4. The label "V_{CC}" represents the positive power supply voltage.

The defining characteristic of the Colpitts oscillator – i.e. what makes it a Colpitts oscillator as opposed to any other type of oscillator – is the way the tank circuit (the parallel tuned circuit) uses a split capacitor to allow the output of the amplifier to be injected across one of the capacitors, while the input to the amplifier is taken from across the other capacitor.

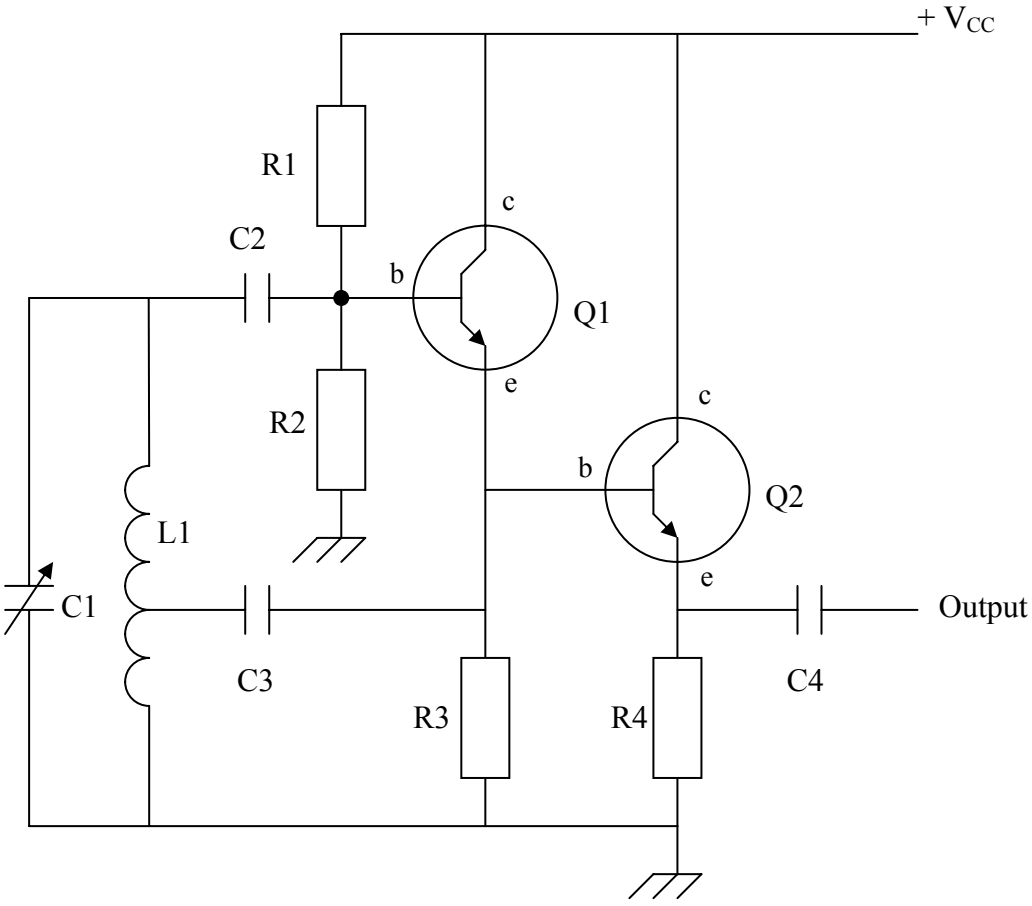
Buffering

Because the amount of signal that is drawn off by the output of the oscillator affects the loop gain of the oscillator, it will also affect the frequency of the oscillator. For this reason it is

important that the amount of signal drawn off does not change, for example in response to a Morse code (CW) transmitter being keyed, otherwise the frequency of the transmitter will change as it is keyed, a phenomenon known as "chirp". Most transmitter designs prevent this by having a *buffer amplifier* between the oscillator and the keyed stages of the transmitter. The buffer amplifier is often a common-collector (emitter follower) amplifier, which shows constant high impedance to the oscillator while having a low output impedance that can supply sufficient current to drive the stages that follow.

The Hartley Oscillator

Another way of feeding the output of the amplifier into a parallel tuned circuit, and the output of the tuned circuit back to the input of the amplifier, is to use a centre-tapped inductor in the tank (tuned) circuit. This is the principal of the Hartley oscillator.



Circuit Diagram of a Hartley Oscillator with a Buffer Amplifier

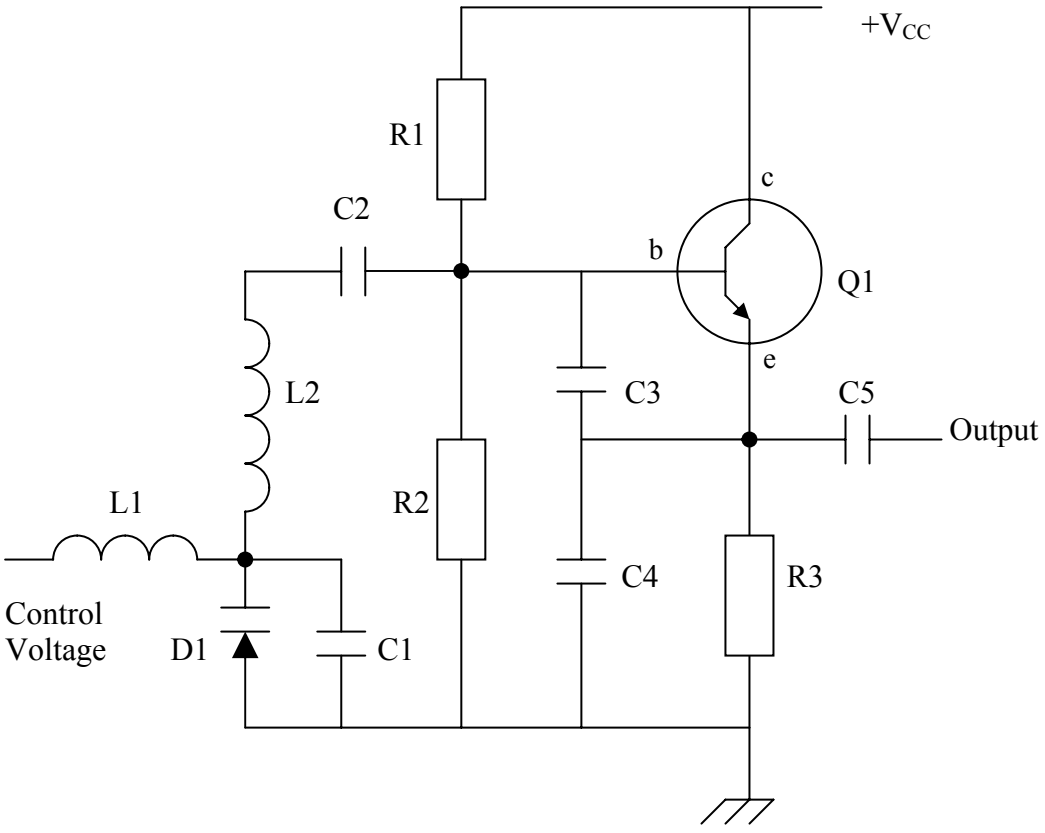
In this circuit, transistor Q1 is a common-collector (emitter follower) amplifier that is biased by R1, R2 and R3. The output of the amplifier, at the emitter of Q1, is coupled via DC blocking capacitor C3 into the parallel tuned tank circuit consisting of L1 and C2 through a tap in the inductor. The tank circuit is coupled back to the input of the amplifier via C2, which serves as another DC blocking capacitor to prevent the base of Q1 from being shorted to earth via L1. The arrow through C1 indicates that it is a variable capacitor, so the resonant frequency of the tank circuit, and hence the oscillator frequency, can be changed by varying C1. The output of the oscillator at the emitter of Q1 is fed to Q2, which is a common-collector

(emitter follower) buffer amplifier. R4 sets the emitter and collector current for Q2. The output of the buffer amplifier is taken from the emitter of Q2 via DC blocking capacitor C4. An Oscillator where the frequency to be varied, typically by turning a control knob, is known as a *Variable Frequency Oscillator* (VFO).

In this circuit, the centre-tapped inductor L1 acts a bit like a step-up transformer, since an AC voltage applied between the centre tap and the chassis connection (the bottom of the inductor) generates a varying magnetic field, which causes a larger voltage to be generated between the "hot" side of L1 (the top of the inductor) and the chassis. This voltage step-up allows the common-collector amplifier to provide power gain in this circuit, despite the fact that the voltage gain between the base and emitter of the transistor is unity (1). A tapped inductor like this is also called an *autotransformer*.

The Voltage-Controlled Oscillator

If part of the capacitance forming the tuned circuit in an oscillator is made up of capacitance from a varicap diode, then the frequency of the oscillator can be varied by changing the reverse-bias voltage applied to the varicap diode. This is called a *voltage-controlled oscillator* (VCO). An example circuit, using a Clapp (series-tuned Colpitts) configuration is shown below:



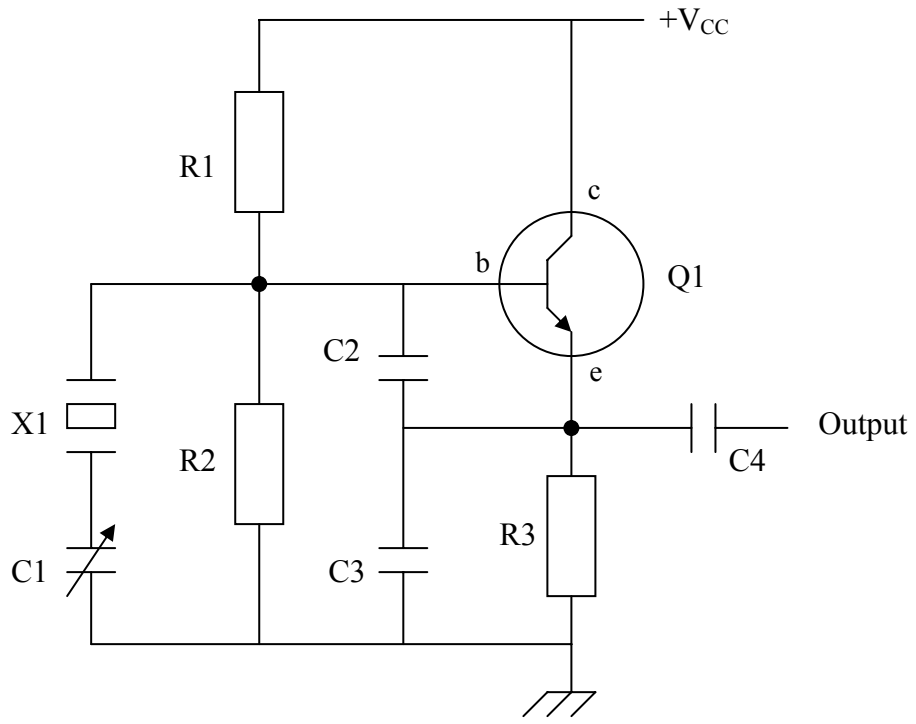
Circuit Diagram of a Voltage Controlled Oscillator

The control voltage is applied through radio-frequency choke L1 to reverse-bias the varicap diode D1. This is in parallel with C1, which provides some additional capacitance (necessary since varicap diodes have fairly low capacitance). They are in series with L2, hence the name "series-tuned Colpitts" oscillator (also called a Clapp oscillator). C2 prevents the DC control voltage from interfering with the bias voltage generated by the voltage divider consisting of

R1 and R2 (or vice-versa). Q1 is operated as a common collector (emitter follower) amplifier, and the output at the emitter of Q1 is fed back into the tank circuit at the junction between C3 and C4, which form the tank circuit along with C1, D1 and L2. The oscillator output is taken from the emitter of Q1 via DC blocking capacitor C5.

The Crystal Oscillator

Quartz crystals exhibit the piezoelectric effect – a voltage applied across the crystal causes the crystal to distort ("bend") slightly, and when the crystal returns to its undistorted shape a voltage is generated across it. As a result, the crystal appears similar to a series tuned circuit and it can be used as the frequency-determining element in an oscillator. A typical circuit is shown below.



Circuit Diagram of a Crystal Oscillator

Here the resonant circuit consists of crystal X1 with series capacitor C1 and capacitors C2 and C3. Q1 operates as a common-collector (emitter-follower) amplifier biased by R1, R2 and R3. The output of the amplifier is fed back into the tank circuit at the junction between C2 and C3. This circuit also uses a "series-tuned Colpitts" or "Clapp" configuration.

Crystals have the advantage of providing very good frequency stability – that is, the frequency of a crystal controlled oscillator will remain stable with little tendency to "drift", which is a problem with oscillators using traditional inductor-capacitor tuned circuits. The disadvantage of crystal oscillators is that they cannot be tuned over any great range. The variable capacitor C1 in this circuit can varying the frequency slightly (which is known as "pulling" the crystal), but the tuning range is very limited. Crystal oscillators that allow the frequency to be varied are called "variable crystal operators", abbreviated "VXO".

Summary

Oscillators are circuits that generate AC signals. Oscillators consist of an amplifier with positive feedback through a phase-shift network. The phase shift network usually also serves as a band-pass filter. An oscillator will oscillate at any frequency and amplitude where the Barkhausen criteria for oscillation are met:

- The loop gain is unity.
- The sum of the phase shifts around the feedback loop is zero or an integer multiple of 360˚.

The output of an oscillator should be buffered to prevent the frequency of the oscillator from changing as the load on the oscillator varies.

There are several different oscillator circuits, including the Colpitts, Hartley and Clapp oscillators, which differ in the precise arrangement of the tank circuit. An oscillator that allows the frequency to be varied is called a Variable Frequency Oscillator (VFO). If the frequency is varied by applying a control voltage, then it is a Voltage Controlled Oscillator (VCO).

Quartz crystals exhibit the piezoelectric effect and act like series tuned circuits. They can be used to control the frequency of an oscillator. Crystal-controlled oscillators exhibit excellent frequency stability, with very little drift. However they are essentially fixed-frequency oscillators; although the frequency can be "pulled" slightly using a variable capacitor, the tuning range is not nearly as wide as for oscillators using ordinary tuned circuits. Crystal oscillators that allow the frequency to be varied are called "variable crystal operators", abbreviated "VXO".

Revision Questions

1 The names Clapp, Colpitts, Hartley refer to:
a. Transistors.
b. Power amplifiers.
c. Oscillators.
d. Diodes.

2 Which of the following is NOT a basic requisite for oscillation?
a. Feedback from output to input of the amplifier.
b. Correct phasing of input and output circuits.
c. Amplifying of signals from input to output.
d. Tuned circuit in both input and output stages.

3 The purpose of an amplifier in an oscillator is to:
a. Cancel phase shift.
b. Compensate for circuit losses.
c. Produce an increasing output.
d. Act as an oscillator buffer.

4 An oscillator varies its frequency as the loading on the following power amplifier is increased. In redesigning this circuit use should be made of:
a. A more powerful oscillator.
b. A well-regulated DC supply.
c. An intermediate buffer stage.
d. Decreased L/C ratio in the oscillator.

5 Colpitts, Clapp, Gouriet, Beat Frequency and Crystal are all types of :
a. Tuners.
b. Oscillators.
c. Antennas.
d. Amplifiers.

6 The characteristic of an oscillator which determines its operating frequency is:
a. Its resistance.
b. Its resonant frequency.
c. Its inductive reactance.
d. Its size.

7 The oscillator configuration where feedback is via a tapped inductor is:
a. The Armstrong oscillator.
b. The Clapp oscillator.
c. The Colpitts oscillator.
d. The Hartley oscillator.

8 A varicap diode might be used in an oscillator to
a. Allow the frequency to be varied by a control voltage.
b. Regulate the supply voltage to the oscillator.
c. Limit the maximum amplitude of the output.
d. Rectify the output waveform to generate an automatic level control voltage.

9 At the frequency of oscillation, the loop gain of an oscillator is:
a. less than 1.
b. exactly 1.
c. greater than 1.
d. zero or an integer multiple of 360.

10 Which amplifier configuration can be used in an oscillator?
a. Common base.
b. Common collector.
c. Common emitter.
d. Any of the above.

Chapter 19 - Frequency Translation

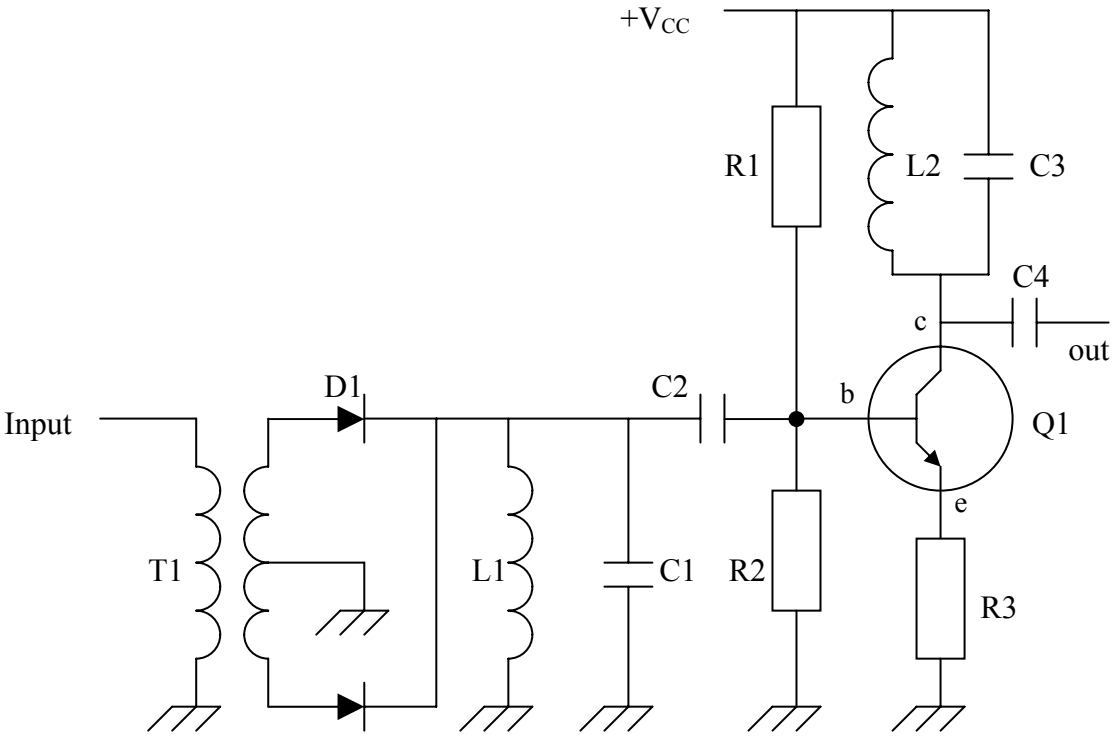
Oscillators are used to generate the signals of various frequencies that are needed by in transmitters and receivers. However often it is useful to be able to create a signal of a desired frequency from signals of other frequencies. For example, it can be very beneficial to generate a signal at the precise output frequency the user has chosen from a very stable reference signal at a fixed frequency. The circuits we use to do this are frequency multipliers, frequency synthesizers and mixers.

The Frequency Multiplier

Any waveform other than a sine wave contains harmonics as well as the fundamental frequency. Harmonics can be found at any integral multiple of the fundamental frequency. For example, a 10 MHz signal that was not a sine wave might have harmonics at 20, 30, 40, 50, 60 and 70 MHz, and so on.

This can be used to create a frequency multiplier. The input sine wave is intentionally distorted; creating a signal that is rich in harmonics. The desired harmonic is then selected using a band-pass filter, yielding a signal that is some integer multiple of the input signal. The most common multiples are 2 and 3. For example, a 7 MHz signal applied to the input of a x2 frequency multiplier would give a 14 MHz signal; applied to a x3 multiplier it would give a 21 MHz signal.

Different types of distortion result in different amounts of the various harmonics. When designing a frequency multiplier, the type of distortion introduced should maximize the desired harmonic. For example, a frequency doubler (a x2 multiplier) could use a full-wave rectifier to distort the input waveform, since the resulting rectified sine wave has a high second-harmonic content. A typical circuit is as follows:



Circuit Diagram of a Frequency Doubler

The input signal is full-wave rectified by T1, D1 and D2. L1 and C1 form a parallel tuned circuit, which is resonant at the output frequency (twice the input frequency). It shorts the DC component of the full-wave rectified signal to ground and attenuates the undesired higher-order harmonics. Transistor Q1 with resistors R1, R2 and R3 form a common-emitter amplifier. There is another parallel tuned circuit made up of L2 and C3 in the collector circuit of the amplifier, which further attenuates undesired high-order harmonics (3, 4, 5 times the input frequency etc.). C2 and C4 are DC blocking capacitors. The output is a sine wave at twice the frequency of the input wave.

A x3 multiplier (frequency tripler) might use a class C amplifier to introduce the necessary distortion, since the output of a class C amplifier has a high third-harmonic component. In VHF and UHF applications, varicap diodes (also known as *varactor diodes*) are often used as the non-linear element to distort the input waveform and generate harmonics.

Because frequency multipliers introduce distortion, they cannot be used with signals that contain a range of frequencies, such as audio signals or amplitude modulated (AM) and single-sideband RF signals. If they were, then the many different frequency components of these signals would interact with each other causing unwanted inter-modulation distortion (IMD) components, that are too close to the desired frequencies to be filtered out. However they can safely be used with un-modulated signals, or with CW (Morse code), frequency modulated (FM) and phase modulated signals.

Frequency multipliers are only useful for multiplying by fairly small numbers, such as 2, 3 or 4. They cannot be used to multiply by large numbers – say 100 – because it would be too difficult to construct a filter to separate the 100th harmonic from the 99th or 101st harmonics, and the nature of frequency multipliers means that they tend to generate at least some amount of most harmonics!

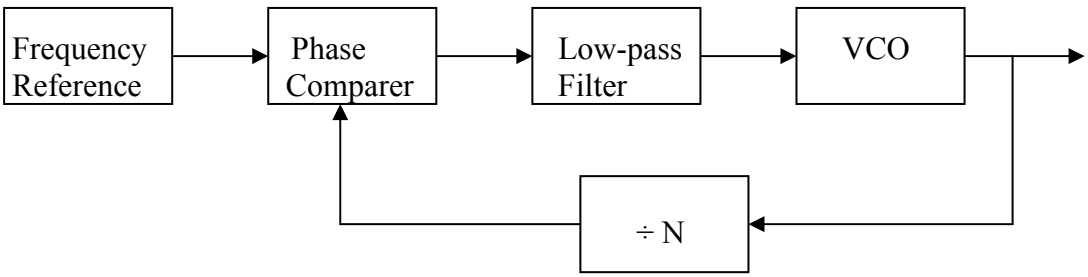
The Frequency Divider

Digital integrated circuits are available that can divide the frequency of an input waveform by any integer number – either a fixed number, or one that can be programmed by a microprocessor. The output of these "digital dividers" is typically a square wave, which contains high harmonic content (especially the odd harmonics at 3 times, 5 times, 7 times the input frequency and so on). These harmonic content can be removed using a suitable low-pass or band-pass filter leaving a sine wave at the desired frequency.

The Phase Locked Loop Frequency Synthesizer

Although variable-frequency oscillators (VFO's) can be used to generate a signal at a frequency selected by the user, they suffer the disadvantage that it is difficult to make them very stable, and their frequency tends to "drift" in response to changes in the ambient temperature, and to make more rapid excursions if bumped or otherwise maltreated. Crystal oscillators, on the other hand, are very stable in the face of temperature variations and mechanical knocks. However their very limited tuning range makes them unsuitable for use as, say, the main oscillator for a transmitter that must cover an entire amateur band.

The most common solution in modern amateur equipment is to use a frequency synthesizer. This is a circuit that can generate many programmable output frequencies based on a single reference frequency derived from a stable crystal oscillator. Although there are several different types of frequency synthesizer, this section will only cover one of these, the phase locked loop (PLL) frequency synthesizer. The block diagram of a simple PLL synthesizer is shown below.



Block Diagram of a PLL Frequency Synthesizer

The output of the frequency reference is fed into a phase comparer. This is a circuit that compares the phase of two signals and generates an output voltage that depends on the phase difference between the signals. This voltage is smoothed by a low-pass filter, and used to control the frequency of a voltage-controlled oscillator. The signal generated by the VCO is the input to a frequency divider that divides the input frequency by some (usually programmable) integer N. The output of the frequency divider is the second input to the phase comparer.

To understand how this circuit works, suppose that the frequency of the VCO is exactly N times the reference frequency. Then the phase comparer will generate a DC output voltage that is dependant on the phase difference between the two signals. This DC voltage will pass through the low-pass filter, and will affect the frequency of the VCO. Suppose the effect is to increase the frequency of the VCO slightly. As the frequency increases, the phase of the VCO output signal will begin to shift relative to the phase of the reference signal, which will change the output voltage of the phase comparer, which is the VCO control voltage.

The circuit is arranged so that if the frequency of the VCO increases slightly, then the resulting output voltage from the phase comparer will reduce the frequency of the VCO again, to bring it back to its “correct” frequency, which is N times the reference frequency. Similarly, if the frequency of the VCO decreases slightly, then the resulting output voltage from the phase comparer will act to increase the frequency of the VCO, again returning it to a frequency N times the reference frequency. In this condition, the VCO is said to be *phase locked* to the reference frequency, since any change in the phase relationship between the two signals (caused, for example, by a change in the VCO frequency) will act on the VCO in a way that will return it to the correct phase relationship with the reference frequency. This is an example of *negative feedback*

In case you were wondering, the reason for the low-pass filter is because most phase comparers actually generate a fairly complex output signal that has a DC (or low-frequency) component that reflects the phase difference between the inputs, as well as components at the different input frequencies to the phase comparer. The low-pass filter rejects the high-frequency outputs, leaving only the low-frequency phase comparison voltage.

So now we have a circuit that can generate a frequency that is N times a stable reference frequency, and is phase locked to the reference frequency, so that it is almost as stable as the reference frequency itself. However by changing the value of N, we can change the output frequency, making it any integer multiple of the reference frequency. If the reference frequency is small enough - say 10 Hz – then we can generate an output frequency that is any multiple of 10 Hz. For example, if the reference frequency is 10 Hz and the divider N is 1 402 000, then the output frequency will be 10 * 1 402 000 = 14 020 000 Hz. If N is increased by 1 to 1 402 001 then the output frequency would be 14 020 010 Hz. This allows us to synthesize almost any desired frequency from a single stable reference frequency. In modern radios, the

divider N that controls the output frequency is usually set by a microprocessor in response to user input, such as adjusting the tuning control.

The only remaining problem is to generate a stable 10 Hz reference frequency for our synthesizer. We can't use a crystal oscillator directly, since 10 Hz is much too low a frequency for a quartz crystal. So what we can do is run a crystal oscillator at a more suitable frequency – perhaps 100 kHz – and then use a digital divider to reduce the frequency to the desired reference frequency. In this case, dividing the 100 kHz oscillator output by a factor of 10 000 would give a 10 Hz reference frequency.

In practical PLL synthesizers it turns out that there is a trade-off between the speed at which the synthesizer can change its frequency (the "tuning rate" if you like) and the resolution of the synthesizer (its "step size"). This is because the resolution of the synthesizer is set by the reference frequency, so a high resolution requires a low reference frequency. But that requires a low cut-off frequency for the low-pass filter, which limits the speed at which the synthesizer can respond to changes in frequency. One solution is to combine the outputs of two synthesizers, one with a high reference frequency that can easily make large frequency changes but has limited resolution, and the other with a small step size that can "fill in" the missing frequencies, but which is never required to make large frequency changes (because the "coarse" synthesizer takes care of that). This is known as a multiple-loop synthesizer.

PLL synthesizers are very versatile and are the basis for most modern transceivers, allowing them to achieve very high stability combined with wide frequency coverage. However, they do have some disadvantages. In particular, early synthesizers such as those found in amateur equipment from the early 1980s suffered from significant phase noise, with the phase and frequency of the output signal varying very slightly as the loop adjusted it to keep it locked to the reference frequency. Modern synthesizer designs are much better in this respect.

The Mixer

Another circuit that is commonly used for frequency translation in both transmitters and receivers is the *mixer*. It is based on the interesting mathematical result that if you multiply two sine waves together, you get a waveform that consists of two components: one with a frequency that is the *sum of* the frequencies of the inputs, the other with a frequency that is the *difference between* the frequencies of the two inputs.

For the mathematically inclined, the relevant mathematical identity is:

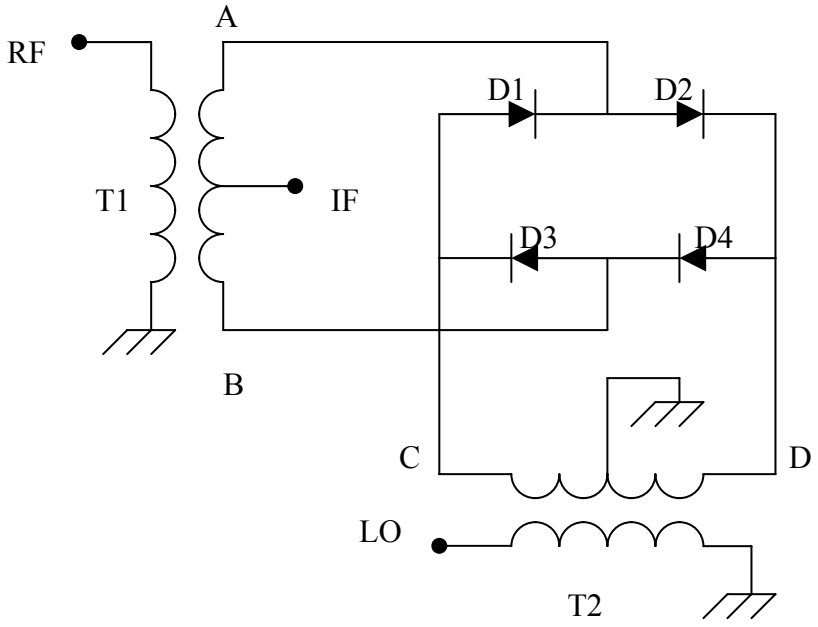
$$2 \sin(A) \cos(B) = \sin(A+B) + \sin(A-B)$$

If you set $A = 2\pi f_1 t$ and $B = 2\pi f_2 t$ then the left-hand side, "2 sin(A) cos(B)" represents two sine waves of frequency f_1 and f_2 with a phase shift of 90˚ multiplied together, while the right-hand side "sin(A+B) + sin(A-B)" represents the superposition (adding together) of sine waves with frequencies $f_1 + f_2$ and $f_1 - f_2$, the sum and difference frequencies. Slightly more complex maths shows that the precise phase difference between the input signals is not important.

For example, if you multiply a 9 MHz sine wave by a 6 MHz sine wave, you end up with two sine waves superimposed: one with a frequency of 15 MHz (the sum of the input frequencies), the other with a frequency of 3 MHz (the difference between the input frequencies). Electronic circuits that do this multiplication are called *mixers*.

Now it turns out that it is not very easy to accurately multiply two sine waves without introducing significant distortion into the output. One common solution is to use a switching mixer. Instead of actually multiplying the two signals together, it uses one of the input signals

to switch the other input signal on and off, or to reverse its direction. Here is a typical circuit diagram for a switching mixer:



A Double-Balanced Diode Mixer

In this mixer, diodes are used as the switching elements. A strong input signal (generally derived from a *local oscillator*) is applied at the point marked LO, while a much weaker radio frequency signal is applied to the point marked RF. The output signal is taken from the point marked IF, for "intermediate frequency". The reason for these names will become apparent once we have studied the design of radio receivers.

The strong LO signal is used to "chop" the weaker RF signal, with the output appearing at the IF port. Here is how it works. Assume that the LO signal's polarity is such that point C is positive with respect to point D. Diodes D1 and D2 will be forward biased (turned on) while diodes D3 and D4 will be reverse biased (turned off). If the diodes are properly balanced, with identical forward bias voltages, then the point between D1 and D2 will be at the same potential as the centre-tap on the secondary winding of T2, that is at chassis (earth) potential. This will earth point A on the secondary winding of T1. If the polarity of the signal applied to the RF port is such that point A is positive with respect to point B, then A will also be positive with respect to the output IF port, so the IF port will be negative with respect to point A, which as we have seen is earthed.

Now suppose the LO signal reverses polarity, while the RF signal remains as it was. Point D is positive with respect to C, so diodes D3 and D4 will conduct, effectively earthing point B. Since the RF signal is making A positive with respect to B, it will also make the IF output positive with respect to B, which is earthed.

So in one half cycle of the LO (switching) input, the RF signal makes the IF output *negative* with respect to earth, while in the other half cycle, the RF signal makes the IF output *positive* with respect to earth. The result is that the LO signal is effectively switching the polarity of the RF signal as it appears at the IF output.

Hold on a moment. We started talking about *multiplying* two signals together, now we are talking about using one signal to switch the polarity of the other. What is the connection? Well it turns out that using the LO signal to switch the polarity of the RF signal is equivalent

to multiplying the RF signal by a square wave with the values +1 and –1. (Multiplied by +1, the polarity is unchanged; multiplied by –1 it is reversed). One effect of this is that, because a square wave contains not only the fundamental frequency, but also many harmonics, these harmonics are effectively mixed with the input signal as well. So instead of only getting the sum and difference frequencies, we also get the sum and difference frequencies of the RF signal and each harmonic of the LO signal. The unwanted mixing products can usually be filtered out by suitable filters following the mixer.

Diode mixers like this one require fairly high drive power at the LO port – typically +7 dBm (5 mW) or more. They usually exhibit a conversion loss of 6-7 dB, meaning that each of the output signals is 6 or 7 dB lower than the RF input signal. However they are widely used in amateur applications because they have good low-distortion properties.

This mixer design is “double balanced” because neither the RF input signal nor the LO input signal will be reflected in the output. An unbalanced mixer would allow both the RF and the LO signals to get into the IF output, while a “single balanced” mixer would allow only one of these signals (typically the weaker and therefore less troublesome RF signal) to make it into the output.

There are many other mixer designs using transistors, specialized integrated circuits and other components.

One big advantage of mixers over other frequency translation circuits (frequency multipliers and the like) is that properly designed mixers do not introduce significant distortion into the signals, and so they can be used with all types of signals, including amplitude modulated (AM), single sideband (SSB) and audio signals.

Summary

Frequency multipliers distort the input waveform to generate harmonics, and then select the desired harmonic using a band-pass filter. They can be used to multiply frequencies by small integers, typically 2 or 3. Frequency multipliers cannot be used with signals that contain many frequencies, such as AM or SSB signals, as they cause too much distortion. However they can be used with CW and FM signals.

Digital integrated circuits are available that can divide a frequency by any integer number.

In a phase locked loop frequency synthesizer, the output frequency is locked to an integer multiple of a stable reference frequency. By changing the multiple, different frequencies can be generated from a single reference frequency. The output of a PLL synthesizer has similar stability to the reference frequency, although it will have additional phase noise.

The output of a mixer will contain signals with frequencies that are the sum of the frequencies of the input signals and the difference between the frequencies of the input signals. Depending on the mixer type, it may also contain signals at the same frequency as one or both of the input frequencies – if both input frequencies are suppressed then the mixer is “double balanced” while if only one input signal is suppressed it is “single balanced”. Switching mixers will also typically contain mixing products caused by mixing various harmonics of the switching (LO) input with the low-level (RF) input. Unwanted mixing products must be removed by suitable filters at the output.

Revision Questions

1 A frequency multiplier stage is generally:
a. Biased into non-linearity.
b. Operated in class A.
c. Used with regeneration.
d. Used in processing SSB signals.

2 The circuit forming the basis of a frequency synthesizer is a :
a. Phase locked loop.
b. Automatic Gain Control.
c. Beat Frequency Oscillator.
d. Power Amplifier.

3 Frequency multiplication is often used in UHF transmitters. This is commonly achieved by applying RF power to diodes and tuned circuits. Such a device is a:
a. Varactor multiplier.
b. Heterodyne mixer.
c. Diode detector.
d. Power amplifier.

4 The reference frequency of a PLL frequency synthesizer is 10 Hz and the programmable divider is set to divide by 315 000. The synthesized frequency will be:
a. 315 kHz.
b. 3,15 MHz.
c. 31,5 MHz.
d. 315 MHz.

5 The cut-off frequency of the low-pass filter in a PLL frequency synthesizer will typically be:
a. Lower than the reference frequency.
b. Higher than the reference frequency.
c. Equal to the output frequency.
d. Higher than the output frequency.

6 A frequency multiplier could be used with the following signal without creating objectionable distortion:
a. An amplitude modulated (AM) signal.
b. A frequency modulated (FM) signal.
c. A single sideband (SSB) signal.
d. An audio-frequency voice signal.

7 A local oscillator signal at 10 MHz is mixed with a 14 MHz signal. The output of the mixer will contain the following frequencies:
a. 10 MHz and 14 MHz only.
b. 4 MHz, 24 MHz and possibly other frequencies as well.
c. 4 MHz and 24 MHz only.
d. 10 MHz, 14 MHz and 24 MHz only.

8 Which of the following circuits can be used to change the frequency of an amplitude modulated signal?

a. A frequency multiplier.
b. A PLL frequency synthesizer.
c. A mixer.
d. Any of the above.

9 As well as the mixing products, the output of a single balanced mixer will contain:

a. Nothing except harmonic mixing products.
b. One of the input signals.
c. Both of the input signals.
d. The average of the two input signals.

10 A switching mixer operates by

a. Reversing the polarity of one of the inputs depending on the polarity of the other.
b. Accurately multiplying two sine waves together.
c. Adding the two input signals together and then distorting the result to generate mixing products.
d. Relying on the square-law transfer characteristic of Field Effect Transistors.

Chapter 20 - Modulation Methods

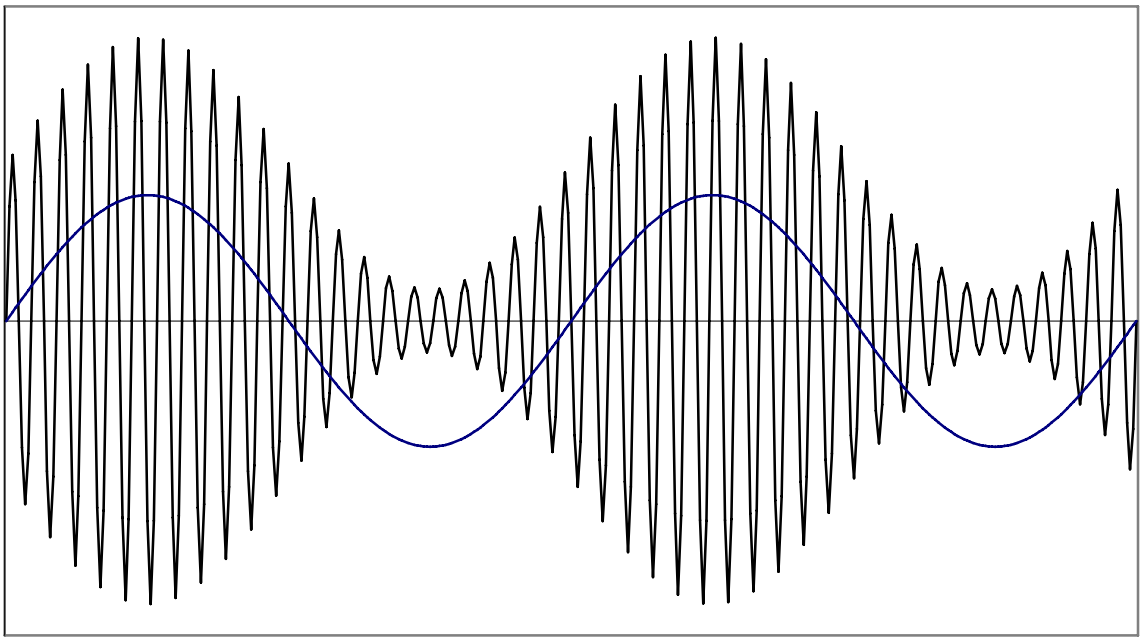
Radio is based on the fact that electromagnetic waves of certain frequencies can travel great distances and still be strong enough to be detected by a radio receiver. However in order for this to be useful, we need a way of sending information with, or imprinted upon, the radio waves. The sort of information that we wish to send – human speech, images or perhaps digital information – is not generally of the correct frequency to benefit directly from the ability of radio to span great distances. For example, the human voice has frequencies that range from approximately 300 Hz to 3 kHz. These frequencies are much too low to be effectively propagated as radio waves.

Modulation is the process of imprinting information on radio waves, so we can take advantage of the propagation of radio waves to transmit the information to a distant receiver.

Amplitude Modulation (AM)

One of the earliest methods of modulation is *amplitude modulation*, or A.M. Although not widely used in the amateur service any more, it still lives on in the A.M. transmissions of commercial radio stations in the medium frequency (or "medium wave") broadcast band. In amplitude modulation, the amplitude (strength) of a radio frequency signal, called the *carrier* is varied according to the amplitude (strength) of the modulating signal.

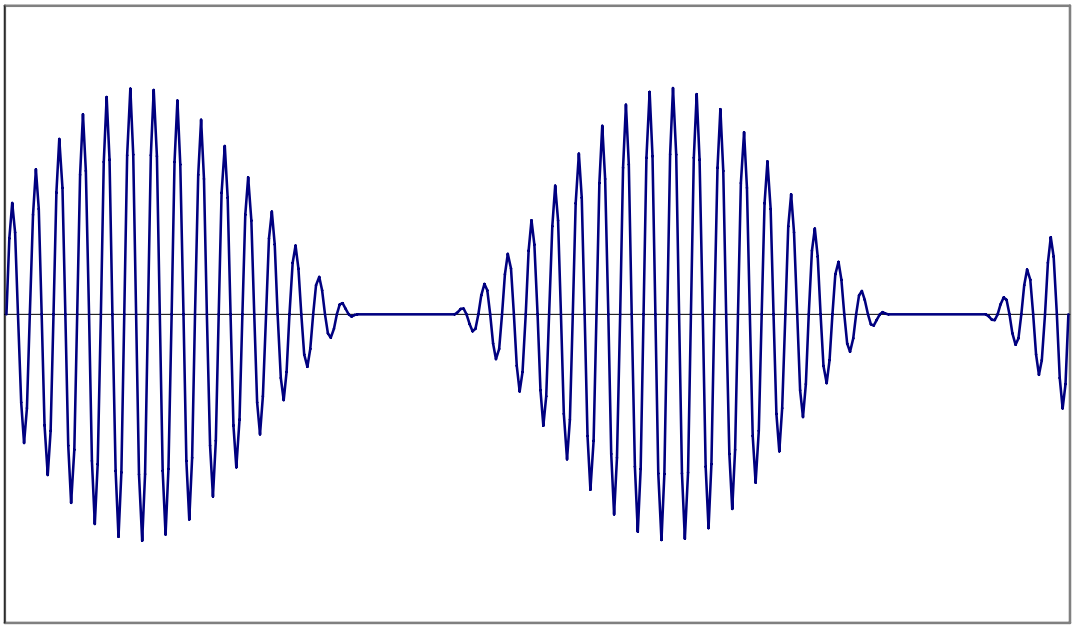
The plot below shows a low frequency sine wave, and the result when this is used to amplitude-modulate a higher frequency carrier.



A modulating signal and amplitude-modulated carrier

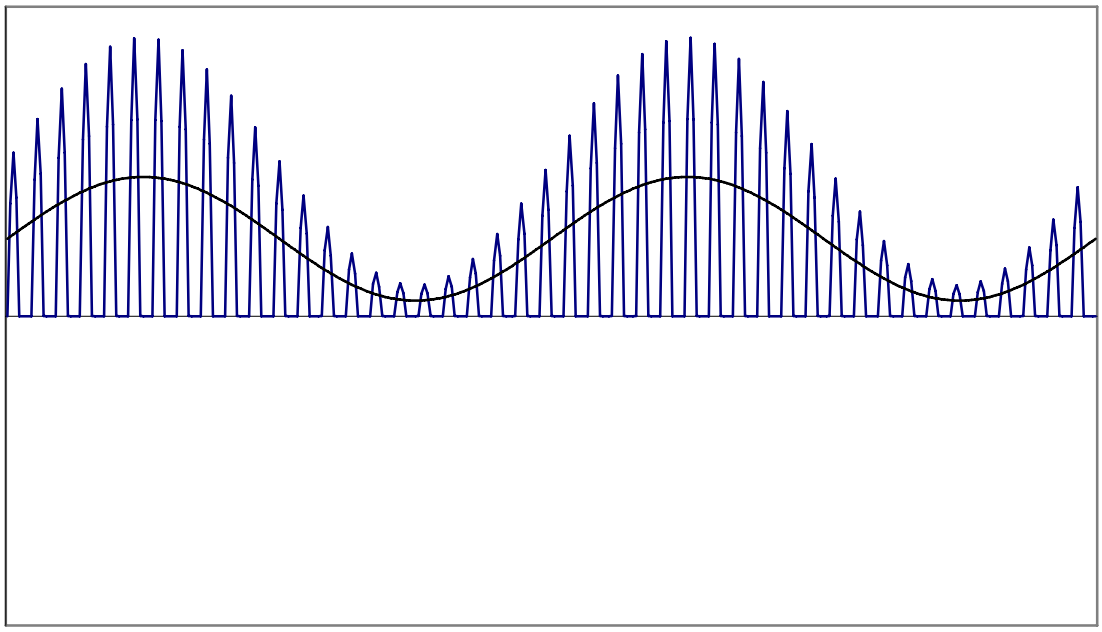
See how the amplitude of the high frequency carrier wave varies in step with the amplitude of the low frequency modulating signal. When the amplitude of the modulating wave is zero, the amplitude-modulated wave is at its "average" output level. When the amplitude of the modulating waveform is positive, the amplitude-modulated signal is above this "average" amplitude, and when the modulating wave is below zero, the output is below this "average" level.

The *modulation depth* of an amplitude-modulated signal is the percentage by which the carrier signals varies above and below its average level in response to the modulating signal. In this example, the carrier is 80% modulated because the peak in the carrier amplitude is 80% above its average level, and the minimum carrier amplitude is 80% below its average level. The maximum possible modulation depth is 100% modulation. In a 100% modulated A.M. signal, the carrier amplitude decreases to zero when the modulating signal is at its most negative. Any attempt to modulate at more that 100% would result in the carrier "bottoming out" at zero amplitude and distorting the modulation signal. This is known as *over modulation* which introduces a great deal of distortion and should be avoided. An example of an over modulated signal is shown below:



An over-modulated A.M. signal

Amplitude modulation has the advantage that it is very simple to recover the modulating signal from the amplitude-modulated signal in the receiver. A simple half-wave rectifier followed by a low-pass filter will recover the modulating signal, which is typically an audio signal. The plot below shows a half-wave rectified A.M. signal, and the result of passing this through a low-pass filter.



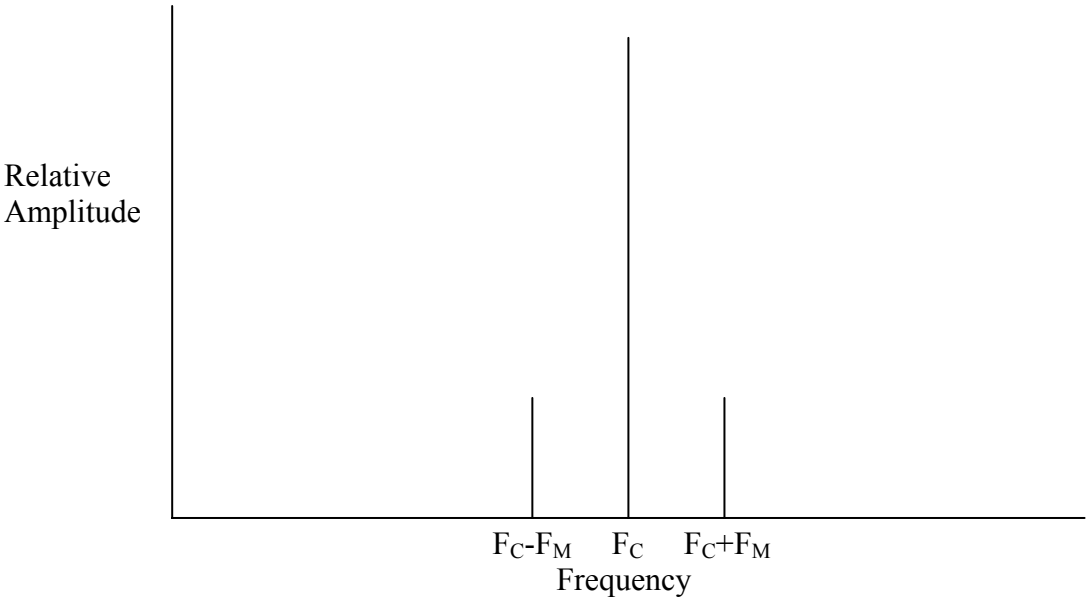
A half-wave rectified A.M. Signal and the recovered modulation

The low-pass filter has removed what remains of the carrier, leaving the modulation "envelope" and a D.C. offset (which is indicated by the fact that the recovered signal is not symmetrical about the X axis). The D.C. offset can be simply removed by a D.C. blocking capacitor to obtain the original modulating signal. The process of recovering the modulation signal from a modulated signal is known as *demodulation*.

Another way of looking at amplitude modulation is that it consists of multiplying the carrier by the modulating signal plus a D.C. offset. The value of the D.C. offset would be chosen to ensure that the sum of the modulating signal and the offset always remains positive, in order to prevent over-modulation. This means that amplitude modulation consists of *mixing* the carrier and modulation signals. Of course we know that mixing two signals results in an output that contains the *sum* and *difference* of the input frequencies, and possibly other components. In this case, the output also includes the carrier wave. This is because the DC offset that we added to the modulating signal has a frequency of zero (because it's D.C.) which also mixes with the carrier, creating a sum frequency (the carrier frequency plus zero) and a difference frequency (the carrier frequency minus zero) that are both the same frequency as the carrier.

So if the carrier frequency is F_C and the modulating frequency F_M, then the amplitude-modulated signal will have frequency components of F_C, F_C-F_M and F_C+F_M. These components can be plotted on a graph that shows frequency on the X-axis, and the relative amplitude of different components of the signal on the Y-axis. This is called the *frequency spectrum* of the signal.

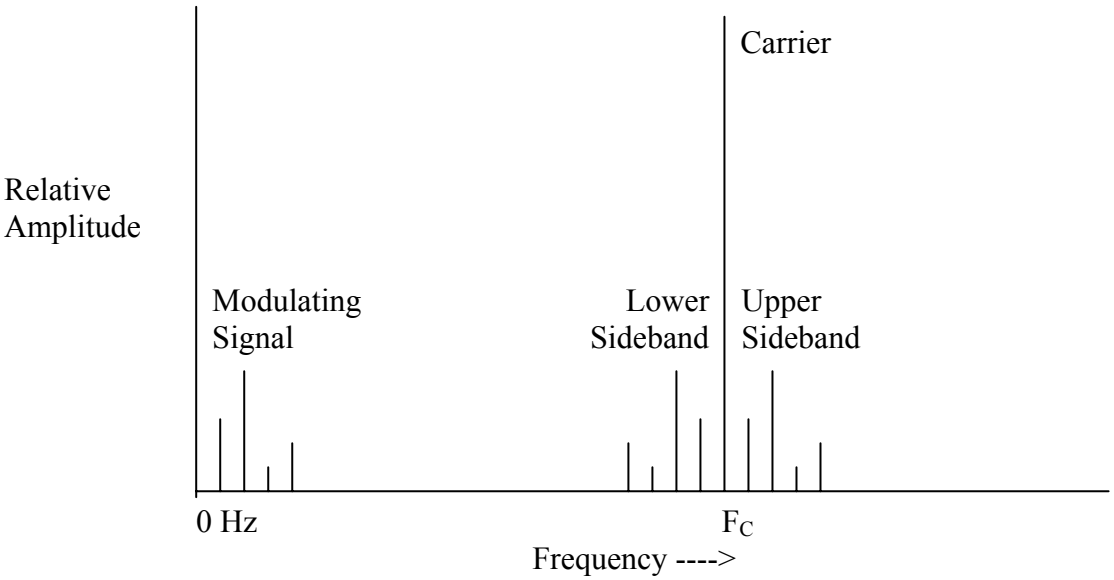
The vertical line above the carrier frequency F_C represents the carrier, while the lines above frequencies F_C+F_M and F_C-F_M represent the sum and difference frequencies respectively. Note that the carrier is much stronger than either of the other components. In an amplitude-modulated signal, two thirds of the power is contained in the carrier; the sum and difference frequencies together make up only one third of the total power of the modulated signal.



The frequency spectrum of a carrier amplitude-modulated by a sine wave

So far we have only considered a carrier that has been modulated by a single sine wave. However speech consists of a whole range of frequencies, with many different frequency components present simultaneously in a speech waveform.

Fortunately it is quite simple to figure out what happens if we amplitude-modulate a carrier with a speech signal that contains many different frequency components. Each of the different frequency components in the speech will create two output signals, one at the sum of the carrier frequency and this component of the modulating signal and one at the difference frequency. The following plot shows the frequency spectrum of some modulating signal (it is on the left of the graph, at a low frequency) and the corresponding amplitude-modulated signal.



Spectrum of a modulating signal and the corresponding amplitude-modulated signal

See how each component of the modulating signal corresponds to two components of the resulting amplitude-modulated signal, on above the carrier (the *sum*) and one below the carrier (the *difference*).

The total of all the "sum" components of the modulated signal – that is, all the components of the modulated signal that are higher in frequency than the carrier – is called the *upper sideband* of the A.M. signal. The total of all the "difference" components – that is, all the components of the modulated signal that are lower in frequency than the carrier – is called the *lower sideband* of the A.M. signal.

In order for speech to be reproduced intelligibly, frequencies from about 300 Hz to 3 kHz are required. This means that for a communications grade A.M. signal, such as is used in the amateur service, the upper sideband will extend from 300 Hz above the carrier to about 3 kHz above the carrier, while the lower sideband will extend from about 300 Hz below the carrier to about 3 kHz below the carrier. So the total *bandwidth* of the signal is 6 kHz, from 3 kHz below the carrier frequency to 3 kHz above the carrier frequency.

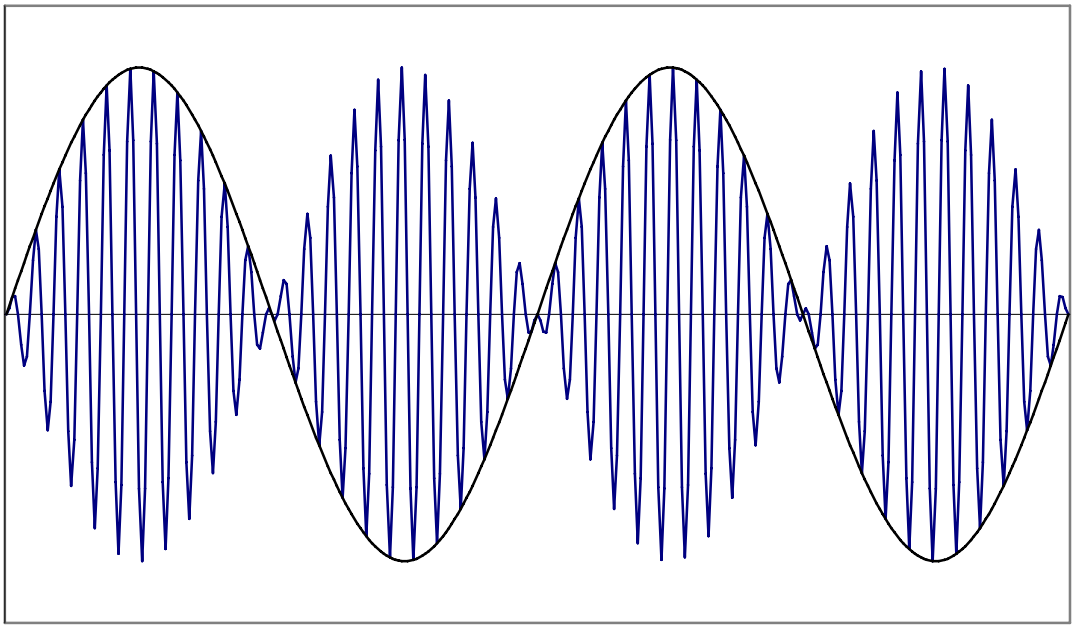
This analysis of the frequency spectrum of an A.M. signal shows the two greatest disadvantages of amplitude modulation.

1. The component of the signal at the carrier frequency conveys no information (it is an unvarying carrier), and yet it consumes two thirds of the power of the signal. This makes amplitude modulation quite inefficient power wise.

2. An A.M. signal transmits two copies of the modulating information, one in the upper sideband and one in the lower sideband, while only one of these would be sufficient to recover the original modulation. This is why the bandwidth of an amplitude-modulated signal is *twice* the bandwidth of the modulating signal, and so A.M. is quite inefficient in terms of the amount of spectrum (frequencies) required. This is particularly important on the crowded amateur bands.

Double-Sideband Suppressed-Carrier Modulation

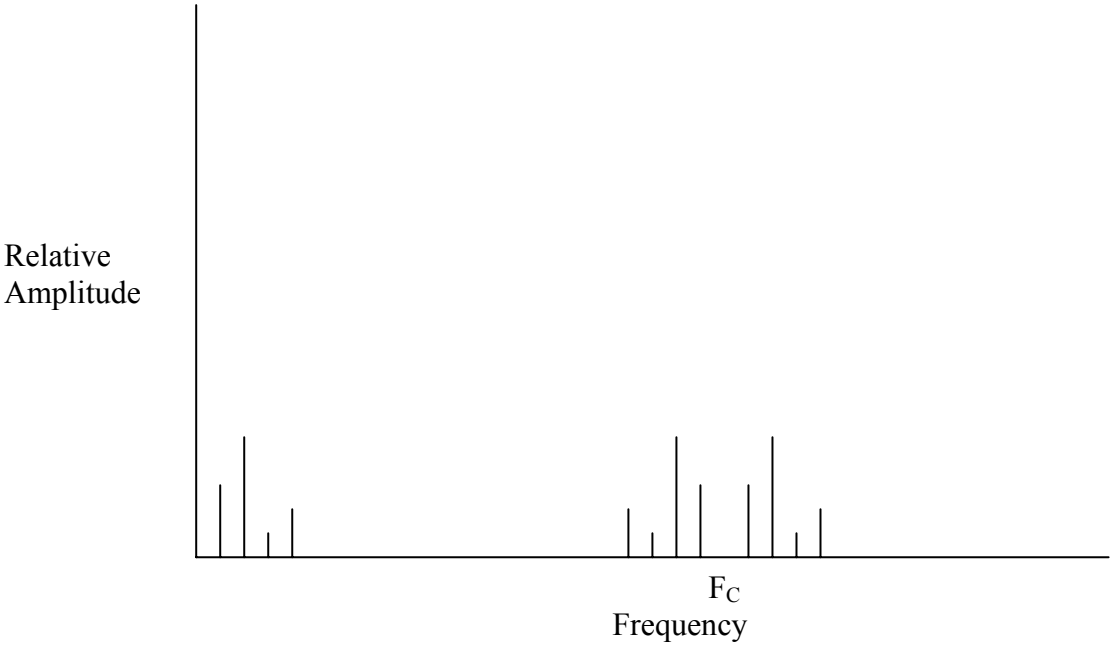
We could overcome the first of these problems – the power wasted by the carrier – if we could generate a signal without a carrier. This can be done by using a *balanced modulator*, which outputs only the sum and difference components, but not the carrier itself. Mathematically this is equivalent to simply multiplying the carrier signal by the modulating signal, without adding any D.C. offset. The plot below shows a low frequency sine wave modulating signal, and the resulting double-sideband suppressed-carrier modulated signal.



A sine wave and double-sideband suppressed-carrier modulated signal

This time, because there is no D.C. offset on the modulating signal, the resulting double sideband modulated signal is zero when the modulating signal is zero. When the modulating signal goes from being positive to being negative or vice-versa, the phase of the modulated signal is inverted, indicating that the modulating signal has crossed the axis. Note that you could not use a simple half-wave rectifier and low-pass filter to recover the modulation.

The frequency spectrum of a double-sideband, suppressed carrier signal is shown below, using the same multi-frequency modulating signal as in the last plot.

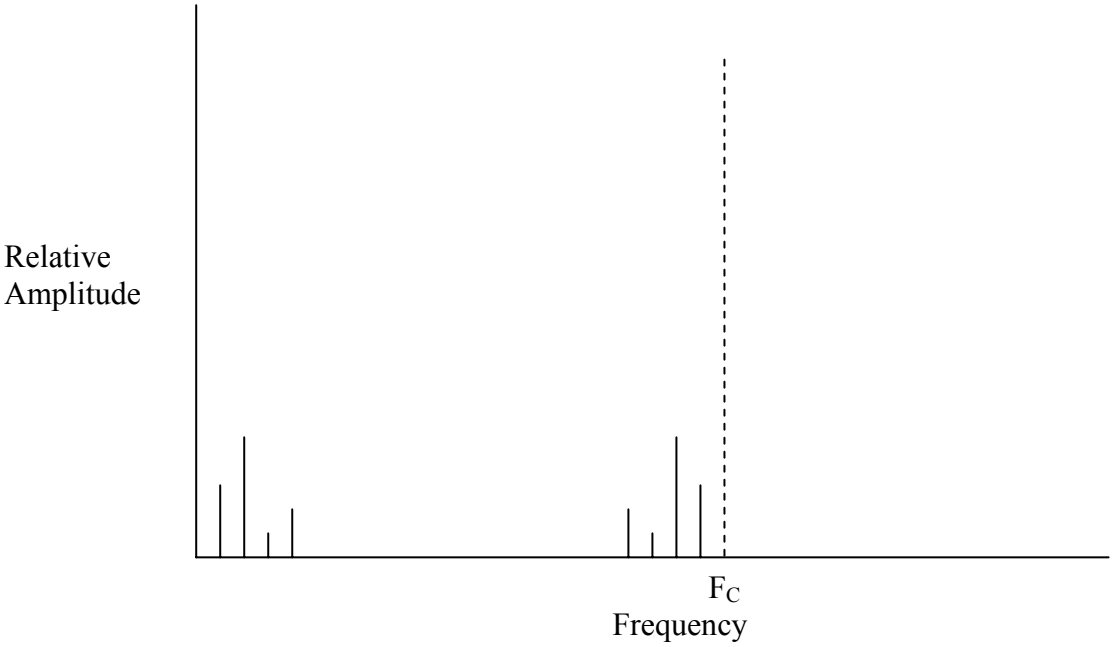


Modulating signal and the corresponding double-sideband suppressed-carrier signal

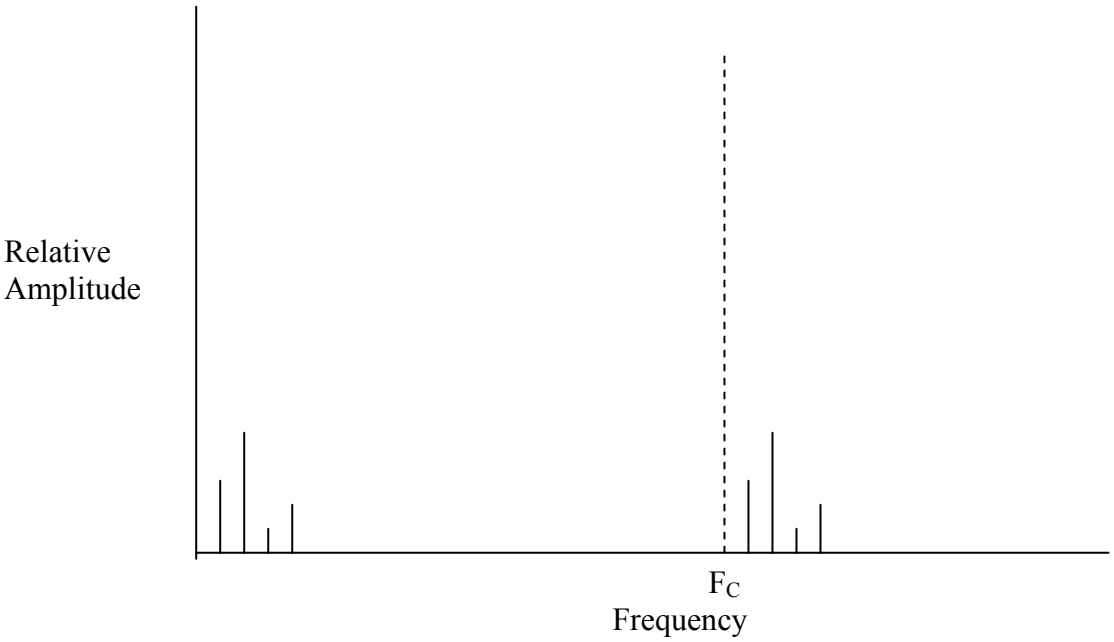
Not surprisingly, it looks exactly like the frequency spectrum of the amplitude-modulated signal, but without the carrier. Double-sideband suppressed-carrier signals are more power-efficient than amplitude-modulated signals, since they do not waste any power on the carrier. However they still occupy twice the bandwidth as the original modulating signal, making them wasteful of spectrum. For this reason, double-sideband suppressed-carrier signals are rarely used in practice.

Single-Sideband (SSB)

In order to avoid wasting bandwidth, we could simply take a double-sideband suppressed-carrier signal and remove one of the sidebands, leaving only a single sideband remaining. This type of modulation is formally known as "single sideband suppressed-carrier modulation", but is usually called just "single sideband" or "SSB". If we remove the lower sideband, then the result would be an upper-sideband (USB) signal. If we remove the upper sideband, then the result would be a lower-sideband (LSB) signal. The two plots below show the frequency spectra have lower-sideband and upper-sideband signals. The carrier frequency is shown as a dotted line so you can see where the frequency spectrum is in relation to where the carrier would have been if it had not been suppressed; but of course the carrier is not actually transmitted.



The frequency spectrum of a modulating signal and the corresponding lower-sideband signal



The frequency spectrum of a modulating signal and the corresponding upper-sideband signal

Note that in the lower-sideband signal, the frequency spectrum of the modulating signal has been inverted (low frequencies in the modulating signal correspond to high frequencies in the lower-sideband signal and vice-versa), while in the upper sideband signal the spectrum in the modulated signal is the same way around as it was in the modulating signal. In fact, an upper sideband signal has an identical frequency spectrum to the original modulating signal, it has just been translated to a higher frequency.

Single sideband is the most commonly used means of transmitting speech in the amateur service. Both upper- and lower-sideband are used. By convention, lower-sideband is used on frequencies below 10 MHz, while upper-sideband is used on frequencies above 10 MHz.

Because SSB signals do not have a carrier, the receiver frequency must be accurately adjusted to properly recover the original audio. Any maladjustment of the receiver frequency will result in the pitch of the audio being slightly too high or too low. This is not important for speech, as it is easy to adjust the receive frequency sufficiently accurately to make speech intelligible, but it is the reason why AM or FM are usually preferred for music transmissions, where even a slight frequency shift in the received audio would be problematic.

Continuous Wave (CW)

Continuous Wave (CW) consists of turning the carrier on and off in order to convey information in Morse code. The name comes from the fact that the first transmitters used sparks, and were not capable of transmitting a continuous signal. Their transmitted signals would consist of an initial strong oscillation when the spark sparked that rapidly died down, known as “damped waves”. So when the first valve-based transmitters became available that were capable of transmitting continuously, they were called “continuous-wave” or “CW” transmitters, despite the fact that information was transmitted by turning the carrier on and off with the resulting dots and dashes standing for letters in Morse code.

It might seem at first as thought he frequency spectrum of a CW transmission should contain only the carrier, since the transmission consists of turning the carrier on and off. However turning a carrier on and off is the same as amplitude-modulating it with a waveform that is at

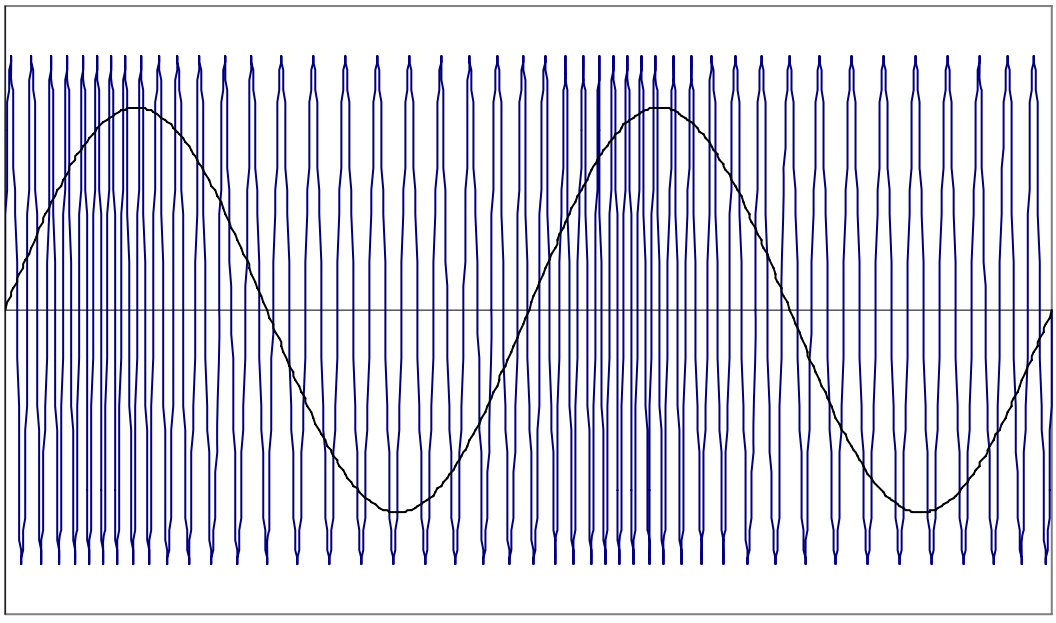
some fixed D.C. level when the carrier is to be turned on, or at zero when it is to be turned off, and so we should expect some sidebands in the keyed signal. These are called “key clicks” because they can be heard as clicks in a receiver when it is tuned close to, but not actually on the same frequency as, as CW transmission.

The shape of the envelope of the CW waveform – that is, the way it is turned on and off – has a big influence on the strength of the key clicks and how far they extend away from the carrier frequency. If the carrier reaches full amplitude as soon as it is turned “on”, and zero amplitude as soon as it is turned “off” then a lot of key clicks will be generated, causing noticeable interference to stations several kilohertz away. To avoid this, the carrier should be allowed to “ramp up” to full volume relatively slowly, and to decay back to zero over a while when it is turned off. The optimum ramp-up and decay period for a CW signal is around 5 ms. This can be achieved using a capacitor that is charged and discharged through a resistor to determine the keying envelope. This acts as a simple low-pass filter, attenuating the high-frequency harmonics of the keying waveform that would otherwise cause key clicks.

Although it may appear archaic, CW is still in widespread use. One of its advantages is that it is intelligible with much lower signal strengths than any voice signal. Practical listening tests have shown that CW requires about 13 dB less power for the same intelligibility as an SSB signal. So a 100 W CW transmitter will “get out” as well as a 2 kW SSB transceiver!

Frequency Modulation (FM)

Instead of varying the *amplitude* of the carrier depending on the amplitude of the modulating signal, frequency modulation (F.M.) varies the *frequency* of the carrier in response to changes in the amplitude of the modulating signal. For example, when the amplitude of the modulating signal is positive, the frequency might be increased slightly from the original carrier frequency, and when the modulating signal is negative, the frequency of the carrier might be reduced slightly. The following plot shows a frequency-modulated signal



A sine wave and corresponding frequency-modulated signal

Note that the amplitude of the signal remains constant, while the frequency varies according to the amplitude of modulating signal. (The amount of frequency change has been exaggerated to make it easier to see.)

The amount that the frequency of the carrier increases or decreases in response to the modulation is called the *deviation* of the signal. The frequency of the carrier is both increased and decreased by the deviation, so for a signal with a deviation of 2,5 kHz, the frequency of the modulated signal will range from 2,5 kHz below the centre frequency to 2,5 kHz above the centre frequency. The centre frequency is the frequency with no modulation applied.

The *deviation ratio* is the maximum deviation divided by the highest modulating frequency. For example, if the deviation is 2,5 kHz and the maximum modulating frequency is 3 kHz then the deviation ratio would be 2500/3000 = 0,83.

The voice-grade FM transmissions typically used by amateurs are referred to as *narrow-band frequency modulation* (NBFM). In NBFM the deviation is kept to about 2,5 kHz and the resulting signal has a bandwidth of 5-6 kHz, comparable to that of a communications-grade AM signal. Commercial FM broadcast stations, by comparison, have a deviation of 75 kHz and a correspondingly much wider bandwidth.

FM signals have the advantage of better audio quality when the strength of the radio signal being received is fairly strong. This is because when an F.M signal is well above the atmospheric noise level, the amplitude variations due to noise have little effect on the receiver, which is only sensitive to variations in the signal frequency and not its amplitude. However the quality of the recovered audio drops rapidly as the signal strength weakens and gets closer to the level of atmospheric noise. For this reason, amateurs mostly use F.M. for local communications in very high frequency (VHF) bands like the 2 m band (144-146 MHz) and ultra high frequency bands like the 70 cm band (430-440 MHz) where signals are usually strong and atmospheric noise is slight. For long-range communications in the high frequency (HF) bands between 3 and 30 MHz, where signals are often weak and atmospheric noise fairly strong, SSB is preferred.

Frequency-Shift Keying (FSK)

So far we have concentrated on “human readable” signals, like the various phone (voice) modes and CW. However an increasing role is being played by digital communications, where radio is used to transmit digital information between two computers. In this case, the information that is being transmitted consists of binary bits (ones and zeros).

A simple modulation method for digital information is frequency-shift keying, where the transmitter transmits one of two possible frequencies depending on whether it is sending a zero or a one. The two frequencies are called the “mark” and “space” frequencies, with the “mark” frequency corresponding to a logic “1” and the “space” frequency corresponding to logic “0”.

FSK is used by modes such as RTTY (radio teletype), which allows interactive communication between two computers and Packet Radio, which provides electronic mail and file transfers over radio links.

Phase-Shift Keying (PSK)

Instead of shifting the *frequency* of the carrier, it is possible instead to shift the *phase* of the carrier depending on whether a one or a zero is being transmitted. The resulting modulation method is called phase-shift keying (PSK). PSK is preferred over FSK in most modern applications because it is more efficient in terms of bandwidth usage.

PSK comes in several different forms. In *binary phase-shift keying* (BPSK), the transmitted signal has one of two different phases, say 0˚ or 180˚, allowing one binary bit (a one or a zero) to be transmitted at a time. In *quad phase-shift keying* (QPSK), the transmitted signal

can have one of four different phases (0˚, 90˚, 180˚ or 270˚), allowing two binary bits to be transmitted at a time.

The most popular amateur mode to use phase- shift keying is PSK-31, which is an interactive digital mode that allows two operators to "chat" to each other in real time over the radio. Everything that either operator types on his or her keyboard is immediately transmitted and displayed on the computer screen of the other operator (and anyone else who is listening). PSK-31 can use either BPSK or QPSK. When using QPSK the increased throughput is used to provide error detection and correction.

Summary

In amplitude modulation (AM), the amplitude of an RF carrier is varied according to the amplitude of the modulating signal. The resulting AM signal consists of the carrier, the upper sideband (at a higher frequency than the carrier) and the lower sideband (at a lower frequency than the carrier). The carrier takes two thirds of the power of an AM signal, with the remaining one-third of the power being shared equally between the upper and lower sidebands. Although AM signals are easy to demodulate using a half-wave rectifier and low-pass filter, they are inefficient both in terms of power (because the carrier conveys no information but takes 2/3 of the power) and bandwidth (since the modulating information is replicated in both sidebands).

A *double-sideband suppressed-carrier* signal is similar to an AM signal but without the carrier. It can be generated using a *balanced modulator*. The resulting signal is more power-efficient than an AM signal, but still uses twice the bandwidth of the modulating signal.

In a *single-sideband suppressed-carrier* (single sideband, or SSB) signal both the carrier and one of the sidebands has been removed, leaving only a single sideband. SSB signals may be *upper sideband* (USB) or *lower sideband* (LSB). In LSB signals the spectrum of the modulating signal is inverted in the modulated signal; in USB, the spectrum is simply translated to a different frequency but is not inverted. SSB is one of the most efficient means of voice communications, especially when signal strengths are low.

Continuous Wave (CW) transmission consists of turning the carrier frequency on or off, and is used to send information in Morse code. CW is effectively a type of amplitude modulation, and the keying sidebands are known as "key clicks". Their extent and strength can be reduced by turning the carrier on and off fairly gradually, over a period of about 5 ms.

In *frequency modulation* (FM) the frequency of the carrier is varied according to the amplitude of the modulating signal while the amplitude remains constant. FM signals are capable of very good audio quality provided the received signal is fairly strong, but quality deteriorates rapidly as the received signal strength weakens. *Narrowband FM* transmissions by amateurs usually have a deviation of 2,5 kHz, resulting in a bandwidth of 5-6 kHz, which is similar to an AM transmission.

Frequency-shift keying (FSK) and *phase-shift keying* (PSK) are used to transmit digital information. In FSK, one of two frequencies is transmitted depending on whether a one or a zero is being sent; while in PSK the phase of the transmitted signal is varied to indicate that a one or a zero is being sent. FSK is used by modes like RTTY and Packet, while PSK is used by PSK-31.

Revision Questions

1 What is the process called which alters the amplitude, phase or frequency of a radio frequency wave for the purpose of conveying information?

a. Alternating.
b. Microphonics.
c. Rectifying.
d. Modulation.

2 The process of extracting information contained in a RF or IF carrier frequency signal is called:

a. Delination.
b. Degeneration.
c. Decoupling.
d. Demodulation.

3 What does suppressing the carrier in an AM signal change the emission type to?

a. Single-sideband suppressed carrier.
b. Double-sideband suppressed carrier.
c. Frequency modulation.
d. Phase modulation.

4 What is one advantage of double-sideband suppressed-carrier transmission over standard full-carrier AM?

a. Only half the bandwidth is required for the same information content.
b. Greater modulation percentage is obtainable with lower distortion.
c. The transmitter is more energy-efficient.
d. Simpler equipment can be used to receive a double-sideband suppressed-carrier signal.

5 A Class C frequency multiplier stage is unsuitable for raising the frequency of an SSB signal because of:

a. Impedance mismatch.
b. Severe distortion.
c. Lack of a carrier.
d. Untuned output circuits.

6 What signal component appears in the center of an amplitude modulated transmitter's emitted bandwidth?

a. The lower sidebands.
b. The subcarrier.
c. The carrier.
d. The pilot tone.

7 In a frequency modulated signal, deviations from the carrier frequency depend on:

a. Amplitude of the audio signal.
b. Ratio of amplitude to frequency of the audio signal.
c. Frequency of the audio signal.
d. Frequency of the original carrier signal.

8 What sideband frequencies will be generated by an am transmitter having a carrier frequency of 7250 kHz when it is modulated less than 100 percent by an 800 Hz pure sine wave?

a. 7250,8 kHz and 7251,6 kHz.
b. 7250,0 kHz and 7250,8 kHz.
c. 7249,2 kHz and 7250,8 kHz.
d. 7248,4 kHz and 7249,2 kHz.

9 The suppression of the carrier wave and one sideband in a transmission is known as:

a. Amplitude Modulation.
b. Frequency Modulation.
c. Single side-band modulation.
d. Double side-band modulation.

10 What determines the bandwidth occupied by each group of sideband frequencies generated by a correctly operating amplitude modulated transmitter?

a. The audio frequencies used to modulate the transmitter.
b. The phase angle between the audio and radio frequencies being mixed.
c. The radio frequencies used in the transmitter's VFO.
d. The CW keying speed.

11 The term Narrow Band FM modulation usually refers to a signal of:

a. +/- 2,5 kHz deviation.
b. 75 kHz deviation.
c. Low power levels.
d. Very stable frequency.

12 The bandwidth of an AM transmission should not exceed:

a. 10 kHz.
b. 20 kHz.
c. Ultrasonic frequencies.
d. 5 to 6 kHz.

13 When the modulation signal reduces the amplitude of modulated wave to zero, this represents:

a. 50 % modulation.
b. 200% modulation.
c. 100% modulation.
d. Overmodulation.

14 The switching on and off by a Morse key of a transmitter to produce different lengths of carrier pulses is called:

a. Current Injection.
b. Keying.
c. Demodulation.
d. Rectification.

15 CW, SSB FM and AM are all types of:

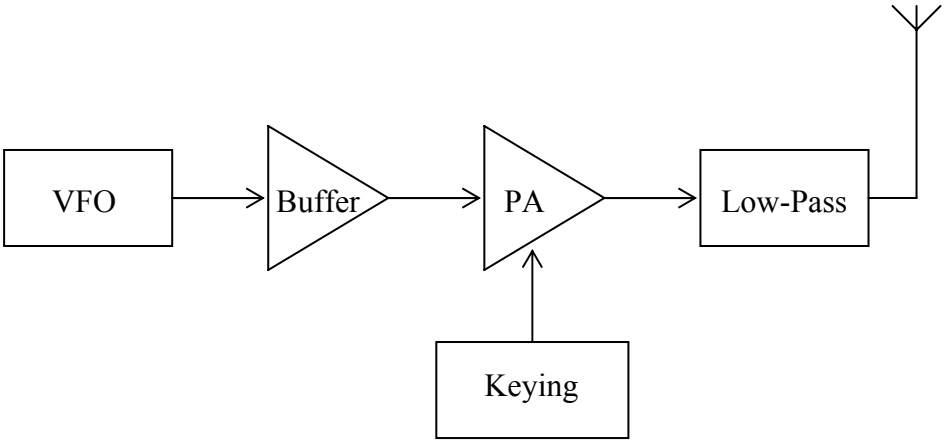
a. Time measurement.
b. Carrier modulation.
c. Radio Waves.
d. Amateur Licenses.

Chapter 21 - The Transmitter

The purpose of a transmitter is to generate a modulated radio-frequency signal that can be applied to an antenna. This module looks at the design of four typical transmitters: a single-band CW transmitter, a VFO-controlled AM transmitter, a simple SSB transceiver and a frequency-synthesized VHF FM transceiver.

A Single-Band CW Transmitter

One of the simplest transmitters is a VFO-controlled single-band CW transmitter. All you need is the variable frequency oscillator, a buffer amplifier (to prevent the variable loading of the power amplifier from affecting the oscillator frequency causing chirp), a keyed power amplifier and a low-pass filter to attenuate harmonics.



A Simple Single-Band CW Transmitter

In this design the PA could run Class C for maximum efficiency since linearity is not required when amplifying a CW signal. This would generate additional harmonics at twice the desired output frequency and higher frequencies, but these could be easily eliminated by the output low-pass filter. The block labeled “keying” should include a keying waveform shaper, to prevent the key-clicks that would be caused by turning the carrier on or off too rapidly.

A design like this would be most suitable for the 80 m (3,5 MHz) or 40 m (7 MHz) bands, to keep the VFO frequency fairly low in order to allow reasonable frequency stability (VFOs are usually best kept below 10 MHz for good stability).

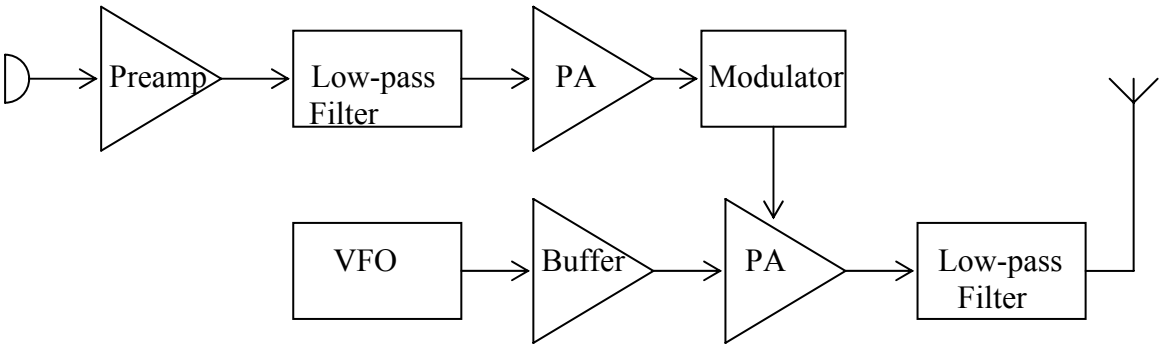
An Amplitude-Modulated (AM) Transmitter

There are two different ways to build AM transmitters. One way is to generate a low-level amplitude-modulated signal, and then amplify this to obtain the desired output power. This has the disadvantage that linear amplification is required because the AM signal contains many frequency components and non-linear amplification would cause inter-modulation distortion. However it is the most common method in modern *multimode* transceivers that can generate AM, SSB and CW signals (and often also FM). This is because low-level modulation is the simplest way to generate an SSB signal, and the same circuitry can also be used to generate an AM signal.

However for specialized AM transmitters there is an alternative, which is to generate the carrier signal and amplify it up to the desired output power, and then use a high-level modulator to modulate it at the full output power. This allows more efficient Class C

amplifiers to be used to amplify the carrier signal, since before it is modulated it contains only a single frequency component (the carrier frequency) and so does not suffer from inter-modulation distortion.

The following circuit shows a VFO-controlled AM transmitter using high-level modulation.

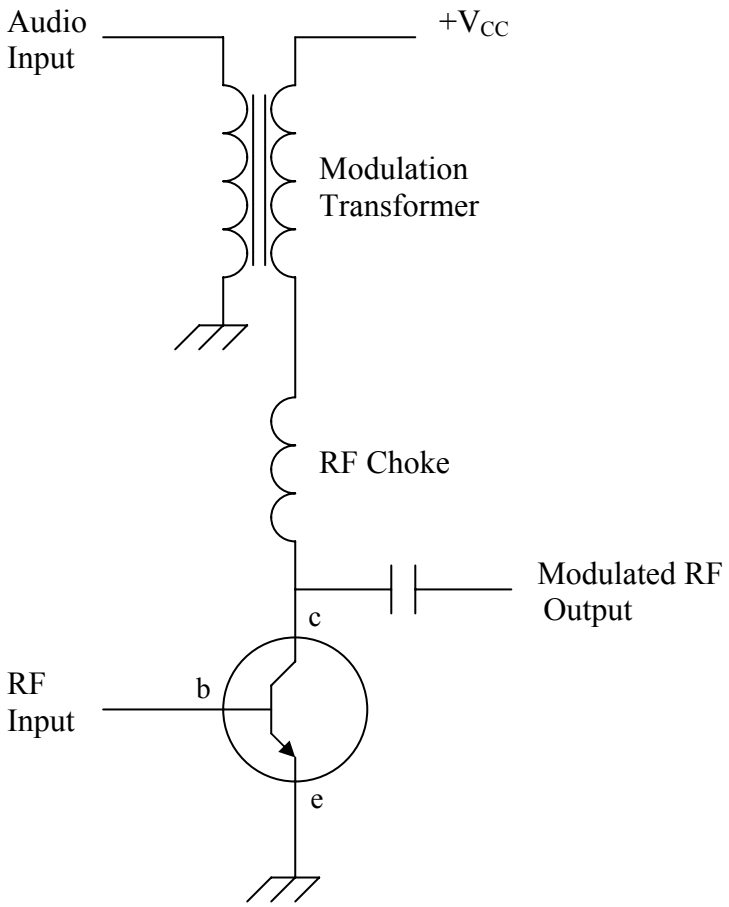


A VFO-controlled AM transmitter using high-level modulation

The audio input from the microphone is pre-amplified and then filtered to remove audio components above the voice range of 300 Hz – 3 kHz. The audio signal is further amplified by a power amplifier and fed to a high-level modulator that controls the Class C RF power amplifier. The input to this amplifier comes from a VFO operating on the intended output signal.

Two-thirds of the energy in an amplitude-modulated signal is contained in the carrier and the remaining one-third in the modulation sidebands. In this circuit, the energy for the modulation sidebands is provided by the audio power amplifier. So if the carrier power were 100 W, then the audio power amplifier would have to supply 50 W to fully modulate the signal.

A high-level modulator typically consists of a *modulation transformer* that modulates the supply voltage to the final output stage depending on the audio modulation. An illustrative circuit diagram is shown below.

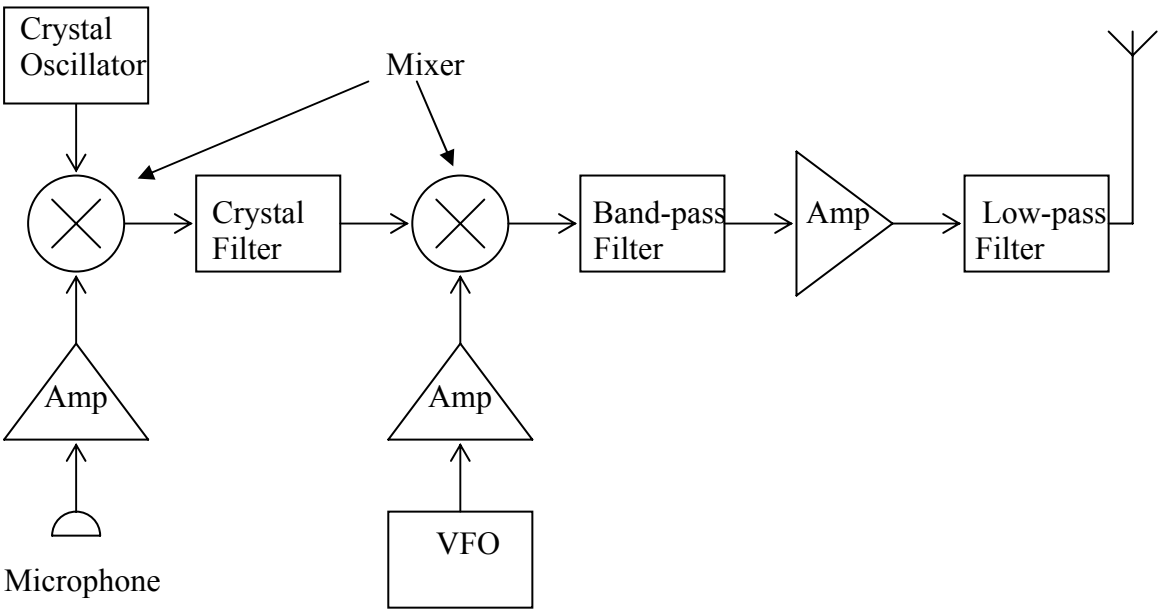


High-level modulation using a modulation transformer

For an example of an AM transmitter using low-level modulation, see the simple SSB transmitter described below. If the balanced modulator is replaced with an unbalanced modulator, and a crystal filter is used that is wide enough to permit both the upper and the lower sidebands to pass, then the result is a low-level modulated AM transmitter.

A Simple SSB Transmitter

The following block diagram shows a simple single-band VFO controlled SSB transmitter for the phone segment of the 20 m band, from 14,100 to 14,350 MHz.



A Simple Single-Sideband Transmitter

In this simple single-sideband transmitter, the carrier is generated by a crystal oscillator at a fixed frequency, perhaps 9,000 0 MHz. This is modulated by the amplified audio input in a balanced modulator (represented here by a circle with a cross inside it, the symbol for a mixer). Because the modulator is balanced, the output signal contains the upper and lower sidebands, but no carrier (so it is a double-sideband suppressed-carrier signal). A very narrow band-pass crystal filter is used to select the upper sideband only, i.e. frequencies from 9,000 3 to 9,003 0 MHz, eliminating the lower sideband. This is called the "filter method" of SSB generation.

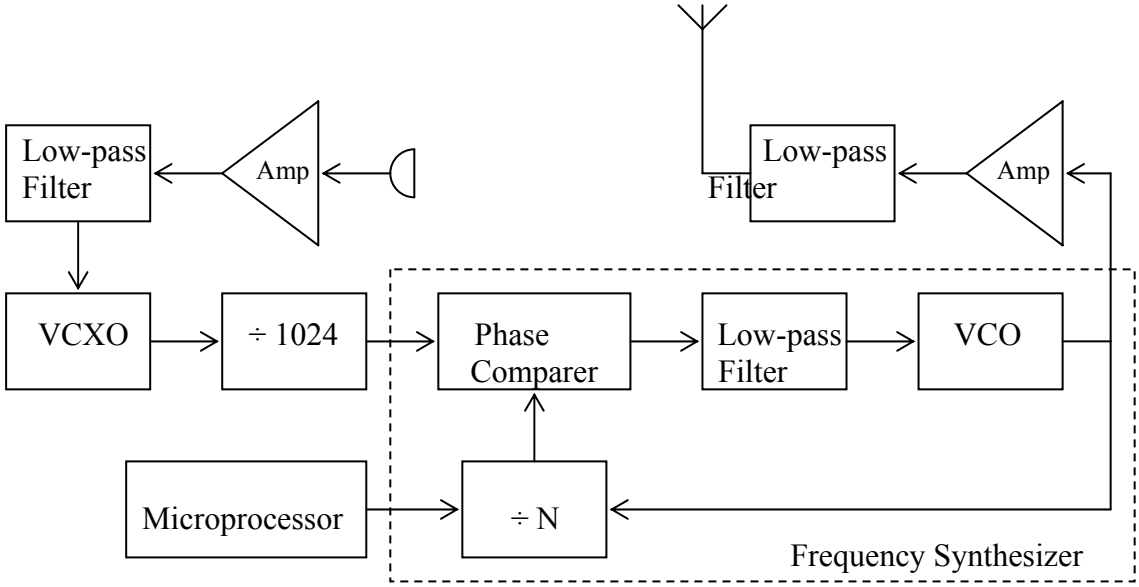
Note that all the filters that are sufficiently selective to pass one sideband while rejecting the other are fixed-tuned, so the resonant frequency cannot be altered. This means that the SSB signal must be generated at a fixed frequency and then mixed up or down to the desired output frequency.

In this case the 9 MHz upper-sideband signal is mixed with the output of a variable-frequency oscillator that ranges from 5,100 to 5,350 MHz, resulting in two signals. The sum will be a USB signal in the range 14,100-14,350 MHz, while the difference will be an USB signal ranging from 3,900-3,650 MHz. The band-pass filter following the mixer is an ordinary inductor-capacitor filter, which is designed to pass the frequency range 14,100–14,350 MHz (the phone segment of the 20 m amateur band) while rejecting frequencies in the range 5,100–5,350 MHz, the unwanted mixing product.

This is followed by a linear RF power amplifier (probably running in class AB) and a final low-pass filter that will pass the desired output frequencies in the range 14,100–14,350 MHz while rejecting harmonics at 28,200 MHz and above.

A Frequency-Synthesized VHF FM Transmitter

Frequency synthesis is a natural approach for building a VHF FM transmitter, since it is not possible to make a VFO run with sufficient stability at VHF frequencies. Also since most FM operation takes place at distinct frequency "channels" spaced 12,5 of 25 kHz apart, a simple single-loop synthesizer will suffice.



A Synthesized VHF FM Transceiver

The signal from the microphone is amplified and then filtered to restrict it to the communications voice range of frequencies below 3 kHz. It is then used to frequency-modulate a voltage-controlled crystal oscillator (VCXO) running at 12,8 MHz, which would probably use a varicap diode to "pull" the crystal frequency slightly. The frequency-modulated output of the crystal oscillator is then divided by 1 024 to generate a frequency-modulated 12,5 kHz reference signal for the PLL frequency synthesizer, which is made up of the functional blocks shown in the dashed rectangle.

Since the voltage controlled oscillator (VCO) in the frequency synthesizer is phase-locked to the reference frequency, it will follow the slight changes to the reference frequency caused by the frequency modulation, so the output of the frequency synthesizer will also be frequency-modulated. To cover the entire 2 m band the "÷N" divider in the frequency synthesizer would range from 11 521 (for an output frequency of 144,012 5 MHz) to 11 679 (for an output frequency of 145,987 5 MHz). The ÷N divider would be controlled by a microprocessor, which would select the correct division ratio according to the frequency set by the user.

The output of the frequency synthesizer would be amplified by the power amplifier and filtered by a low-pass filter to remove harmonics.

Chapter 22 - Receiver Fundamentals

A radio receiver is the heart of any amateur radio installation, whether it is a stand-alone receiver or combined with a transmitter as a *transceiver*. It is relatively easy to build a good transmitter – all you really need is good frequency stability, adequate power and a clean output signal (no harmonics, key clicks or inter-modulation distortion). It is much harder to build a good receiver, and consequently there is more variation in receiver capability amongst both commercial and homebuilt designs.

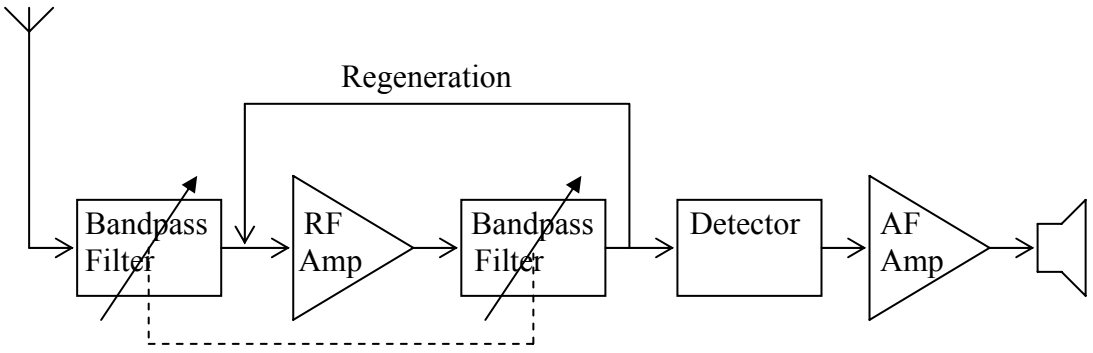
When conditions are good (i.e. radio signals are propagating long distances) the amateur bands can be a very crowded place. If you listen during any CW contest, for instance, you will hear signals spaced 200 to 300 Hz apart over the entire CW section of a band. So the first attribute a good receiver must have is *selectivity*, the ability to distinguish between close-spaced signals and receive only the one that the listener is interested in. Many of the signals on amateur bands are very weak, having come from low-powered transmitters a long distance away, so the second attribute an amateur receiver needs is *sensitivity*, the ability to "hear" very weak signals. And since these weak signals may be adjacent to strong signals, perhaps from other amateurs in your town, amateur receivers need another attribute: *dynamic range*. Dynamic range is the ability of the receiver to receive signals of widely different signal strengths. This is usually achieved by the receiver self-adjusting sensitivity for widely differing signal strengths. This is usually done through the AGC mechanism. Where the AGC has insufficient range, this can be supplemented by an input attenuator or by manual adjustment of the RF Gain.

To get an idea of the challenges faced by receiver designers, a typical weak signal on an amateur band might deliver a power of –120 dBm from the antenna – that's one billionth of a microwatt. A strong signal might deliver –30 dBm, or 1 μW. So a strong signal could be 90 dB (one billion times) as strong as a weak signal – and yet the receiver might need to select and amplify the weak signal to a usable level, without being affected by the strong signal a few kilohertz away!

This module introduces two simple receiver designs – the *tuned radio frequency* receiver and the *direct-conversion* receiver, and considers how well they meet these requirements. It also introduces many of the concepts that you will need for the next module, which covers the *superheterodyne receiver*.

The Tuned Radio Frequency (TRF) Receiver

One of the simplest receiver designs, which has been with us almost since the dawn of radio, is the *tuned radio frequency* receiver. The principle is simple: you use a band-pass filter to select the signal you want, amplify the weak radio signal, demodulate the signal (to recover the audio modulating frequency) and then amplify the recovered audio sufficiently to make it audible in headphones or a loudspeaker. The block diagram below shows the layout of a TRF receiver. The block labeled "detector" is a half-wave rectifier to demodulate AM signals,



A Tuned Radio Frequency Receiver with Regeneration

The arrows through the bandpass filter indicate that they are tunable, so they can be used to select the desired signal. The dotted line joining the arrows on the two bandpass filters mean that they tune *together*, so a single control will change the tuning of both filters together.

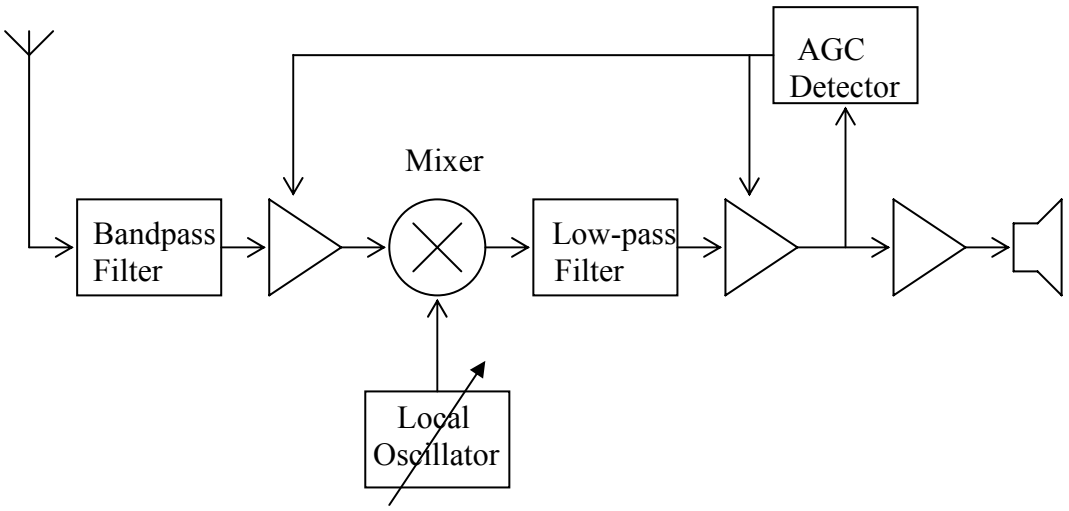
Many TRF receivers use *regeneration*, which means feeding some of the signal from the output of the RF amplifier back to its input, in such a way as to reinforce the signal at the input of the RF amplifier. This is a form of *positive feedback*. It has the benefit of increasing the amplification of the RF amplifier (because some of the signal "circulates" through it many times, being amplified each time) and also increasing the selectivity, since the signal also passes through the band-pass filter at the output of the RF amplifier many times. Of course an amplifier with positive feedback is an oscillator, so if too much regeneration is applied then the circuit will oscillate. Regenerative receivers (a name for TRF receivers that use regeneration) usually have a control to adjust the amount of regeneration, which is adjusted to get the maximum possible sensitivity and selectivity without oscillation.

The advantage of TRF receivers is that they are simple to construct and require relatively few components – typically just two or three valves or transistors and a handful of other parts. This made them attractive in the days before transistors, when thermionic valves were used for amplification in radio receivers, as valves were relatively expensive so the fewer the better!

Their big disadvantage is that they have very poor selectivity and dynamic range. Tunable bandpass filters just aren't capable of rejecting an unwanted signal that is only a couple of kilohertz away from the signal you are listening to, so unwanted signals will also get through to the detector and be recovered as audio or cause inter-modulation distortion. TRF receivers are also best suited for receiving AM signals. Although regenerative receivers can be used with CW and SSB signals, by adjusting the regeneration control so the circuit just oscillates, adjustment is tricky and the quality of reception poor. For these reasons TRF receivers are not widely used anymore.

The Direct-Conversion Receiver

A design that is used in quite a few homebuilt receivers is the Direct Conversion receiver. In a Direct Conversion receiver, the radio-frequency signal from the antenna is mixed with a locally generated oscillator signal, producing the usual sum and difference mixing products. The frequency of the oscillator that generates this local mixing signal – it is known as the *local oscillator* (LO) or *beat frequency oscillator* (BFO) – is set so the difference mixing product is at audio frequency. In this way the Direct Conversion receiver "directly converts" the desired radio-frequency signal to audio, where it can be filtered and amplified. Let's look at the circuit in a little more detail.

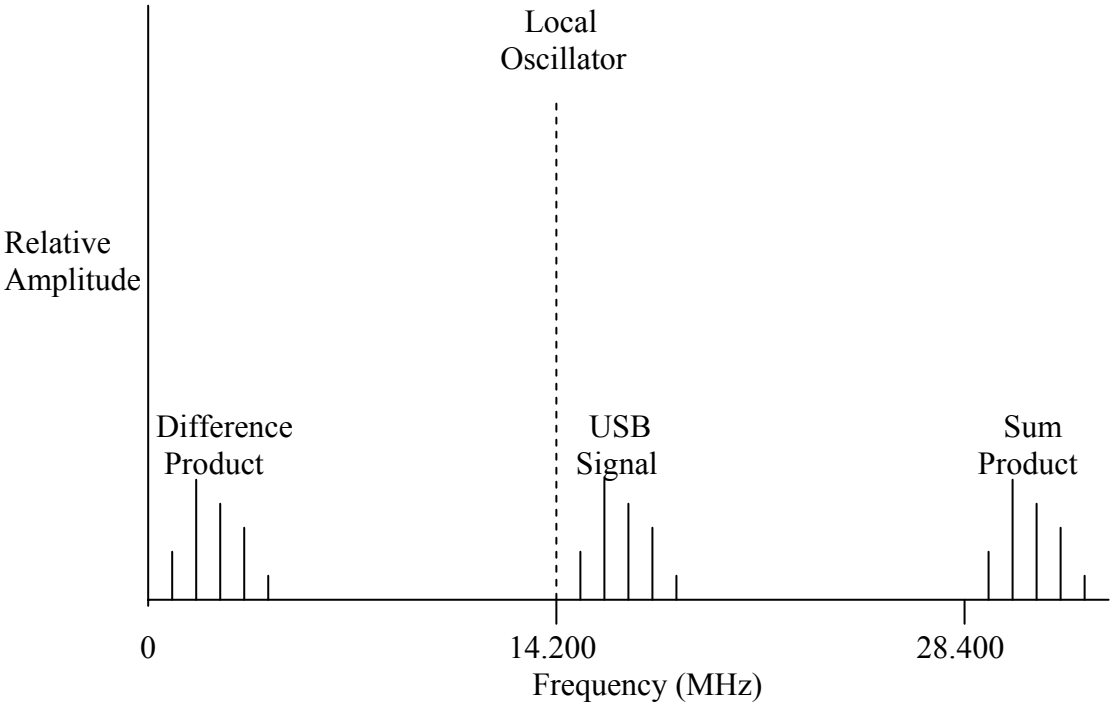


A Direct-Conversion Receiver

The signal from the antenna first passes through a bandpass filter. Unlike in the Tuned Radio Frequency receiver, this bandpass filter is not responsible for the overall selectivity of the receiver – its role is simply to reject interference from strong local commercial broadcast stations and the like. It does not have to be tunable – usually a fixed-tuned filter covering an entire amateur band will suffice.

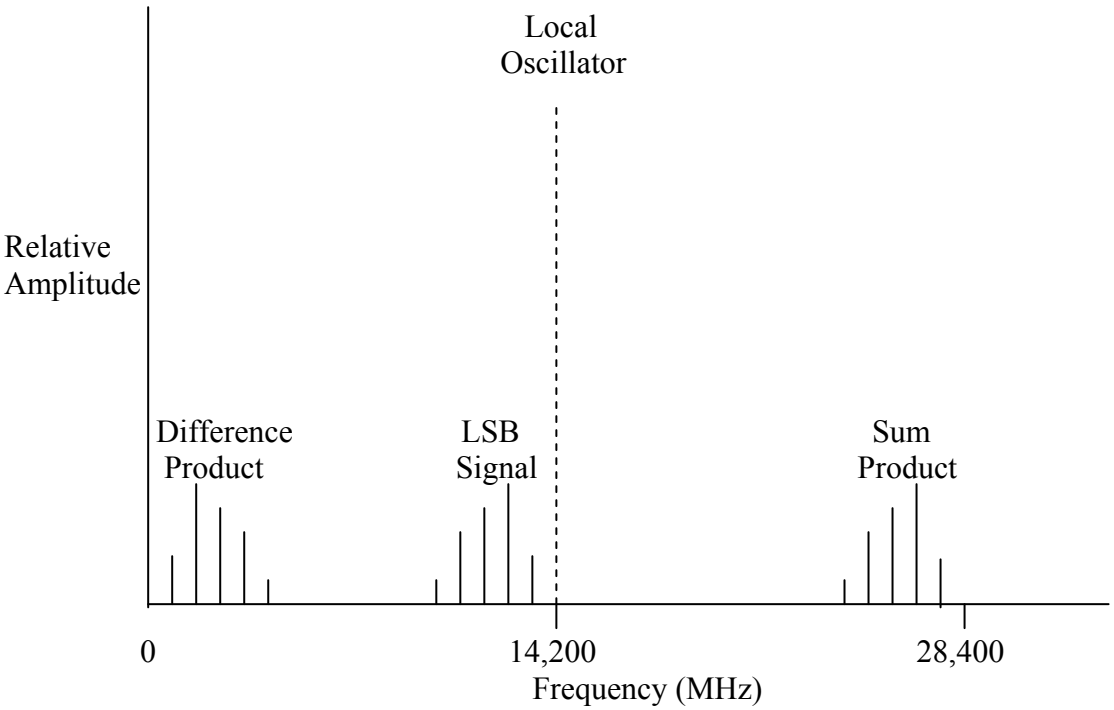
The signal is then amplified by an RF amplifier and fed into the product detector, which We have represented on the diagram using the symbol for a mixer – the circle with a cross in it. ("Mixer", "Modulator" and "Product Detector" are different names for essentially the same circuit, depending on the exact role it plays.) The product detector mixes the amplified RF signal with a signal generated by the tunable local oscillator, generating the usual sum and difference mixing product.

Suppose we want to receive an upper-sideband signal on 14,200 MHz. By convention, we refer to the frequency of a single-sideband signal as the frequency where the carrier would have been if it had not been suppressed. So the upper sideband of this USB signal (i.e. all that is left of it after the carrier and lower sideband were removed) will range in frequency from 14,200 3 MHz to 14,203 0 MHz, 300 Hz to 3 kHz above the (suppressed) carrier. If the local oscillator is set to exactly 14,200 MHz – the frequency where the carrier would have been – then the difference mixing products will range in frequency between 300 Hz and 3 kHz. What we have done is to translate the USB signal from its frequency of 14,200 MHz back to the audio frequency range.



This graph shows how mixing the 14,200 MHz USB signal with a 14,200 MHz signal from the local oscillator generated a difference mixing product (signal frequency – local oscillator frequency) in the audio range and a sum product (signal frequency + local oscillator frequency) up above 28,400 MHz.

Although the example used an upper sideband signal, the same process would work equally well using a lower sideband signal, and the local oscillator frequency would still be 14,200 MHz, the frequency where the carrier would have been. The following graph shows the same process with a lower-sideband signal.



Once again the difference product is back at audio frequency, while the sum product is at around twice the signal frequency, 28,400 MHz. Also note how for the lower side-band signal, the mixing process has inverted the sideband (so the recovered audio is the mirror image of the sideband), which makes up for the sideband inversion that would have occurred when the LSB signal was generated.

So whether the signal is USB or LSB, mixing it with a local oscillator with the same frequency that the carrier would have had will demodulate it and recover the audio.

To complete the hat trick, suppose we have a CW signal at 14,200 MHz. All we need to do is set the local oscillator just below it – say at 14,199 4 MHz, which is 600 Hz below the CW signal – and the difference mixing product will be a 600 Hz tone, just right for listening to CW. So we can also use the product detector to receive a CW signal. (Setting the local oscillator 600 Hz above the CW signal would work just as well.)

We now pass the recovered audio through a low-pass filter. The main purpose of the filter is to remove the difference mixing product from signals near to the one that we are listening to. For example, suppose there is a CW signal at 14,205 MHz while we are listening to our 14,200 MHz USB signal. The difference mixing product of the 14,205 MHz CW signal and the 14,200 MHz local oscillator is 5 kHz – in other words, we have translated the (unwanted) CW signal downwards in frequency to the audio range just as we have translated the (wanted) USB signal to audio. However a low-pass filter with a cutoff frequency of around 3 kHz or so should be able to remove the unwanted CW signal without affecting the desired USB signal.

Because it is quite easy for a strong signal to overload a mixer, causing inter-modulation distortion, the gain ahead of the mixer (i.e. the gain of the RF amplifier) is usually kept quite low so as not to amplify unwanted strong signals and overload the mixer. This means that most of the gain in a Direct Conversion receiver is at audio frequencies, in the amplifiers following the low-pass filter.

The only remaining part of the circuit is the Automatic Gain Control (AGC) system. Because there is such a wide range of signal strengths on the amateur (and other) bands, it is useful to have some way of automatically controlling the gain of the receiver, so it can have a lot of gain to amplify weak signals, but reduce this gain to avoid overload when amplifying strong signals. While this could be achieved with a manually operated gain control, this is not very operator friendly because when tuning from a weak signal (with the gain set on full) to a strong signal, the strong signal can be painfully loud. And when tuning from a strong signal (with the gain turned right down) to a weak signal, you might miss the weak signal altogether unless you remembered to turn the gain up.

The solution is automatic gain control. The AGC detector samples the audio signal after the first audio amplifier, and automatically adjusts the gain of the RF amplifier and the audio amplifier to keep the output signal level fairly constant. The output signal is then amplified by a final audio power amplifier and used to drive headphones or a speaker. The AGC control voltage is often also used to drive a *signal strength meter*, known as an “S meter”, that indicates the strength of the received signal using a fairly arbitrary scale calibrated from S1 (a very weak signal) to S9 (a very strong signal).

The Direct Conversion receiver has several advantages over a TRF receiver. Most importantly, its selectivity is very good, because unwanted nearby signals are easily filtered out by the audio low-pass filter that follows the product detector. It is more stable, having no tendency to oscillate like regenerative TRF receivers do. And it is easy to receive single sideband and CW signals with a Direct Conversion receiver – you just tune the signal in, without having to fiddle with the regeneration control.

However the Direct Conversion receiver does have one significant disadvantage. Since the same local oscillator frequency can be used to tune either a upper sideband or a lower sideband signal, if you are listening to say an upper sideband signal and there is a different signal occupying the frequencies on the other side of the local oscillator where the lower sideband would have been, then the other signal will also be shifted to audio frequencies and will interfere with the station you are trying to listen to.

For example, suppose you are listening to an USB signal at 14,200 MHz as before, but there is also a CW signal at a frequency of 14,199 MHz. Mixing the 14,200 MHz local oscillator signal with the 14,199 MHz CW signal will generate a 1 kHz audio tone. Since this falls within the same 300 Hz – 3 kHz audio range as the desired USB signal, you cannot filter it out using the low-pass filter. And because the unwanted signal is so close in frequency to the desired signal, you can't use the RF bandpass filter to reject it either.

The unwanted signal on the other side of the local oscillator signal is called an "image", so the principal disadvantage of the Direct Conversion receiver can be described as its inability to reject images, or lack of "image rejection". There are more sophisticated variations of the basic Direct Conversion design that *are* able to reject images, but these are quite complex and fall outside the scope of this course.

Summary

The key attributes of a receiver are sensitivity, selectivity and dynamic range. Sensitivity is the ability to receive weak signals; selectivity is the ability to distinguish between nearby signals; and dynamic range is the ability of the receiver to receive signals of widely different signal strengths.

In the tuned radio frequency receiver all signal filtering is done at radio frequencies. As a result they have poor selectivity. Regeneration, which consists of feeding some of the output signal back to the input of the RF amplifier, can increase both the sensitivity and selectivity of the TRF receiver, but makes it prone to oscillation.

In the *direct-conversion* receiver, the incoming RF signal is mixed down to audio frequency using a product detector and local oscillator. Most of the selectivity of a direct conversion receiver is contributed by audio filters following the product detector. Direct conversion receivers have much better selectivity than TRF receivers, but they suffer from an image response to the opposite sideband that can only be eliminated with complex designs.

Revision Questions

1 The specification "1 µV to provide better than 20 dB signal-plus-noise to noise ratio in a passband of less than 1 kHz" would refer to:

a. Sensitivity.
b. Selectivity.
c. Stability.
d. Image rejection.

2 The ability of a receiver to extract weak signals and amplify them to a readable level is known as the receivers':

a. Sensitivity.
b. Selectability.
c. Q factor.
d. Gain factor.

3 The sensitivity of a communications receiver can best be varied by:

a. Altering the input voltage.
b. Altering the RF gain.
c. Changing the IF frequency.
d. Adjusting the volume control.

4 The dynamic range of a receiver can be described as:
a. Its Audio output.
b. The tuning range.
c. The operating voltage.
d. The range of signals over which it operates satisfactorily.

5 The RF stage of a receiver is used to:
a. Improve its sensitivity.
b. Improve its selectivity.
c. Change the frequency.
d. Change the signal tone.

6 The ability of a receiver to receive the desired signal whilst rejecting other frequencies is known as:
a. Sensitivity.
b. Selectivity.
c. A Tuning scale.
d. Wavelength.

7 The circuit that lowers a radio receiver's gain as the received signal becomes stronger is known as:
a. AGC.
b. Filter.
c. Smoothing choke.
d. Selector.

8 What is an S-meter?
a. A meter used to measure sideband suppression.
b. A meter used to measure spurious emissions from a transmitter.
c. A meter used to measure relative signal strength in a receiver.
d. A meter used to measure solar flux.

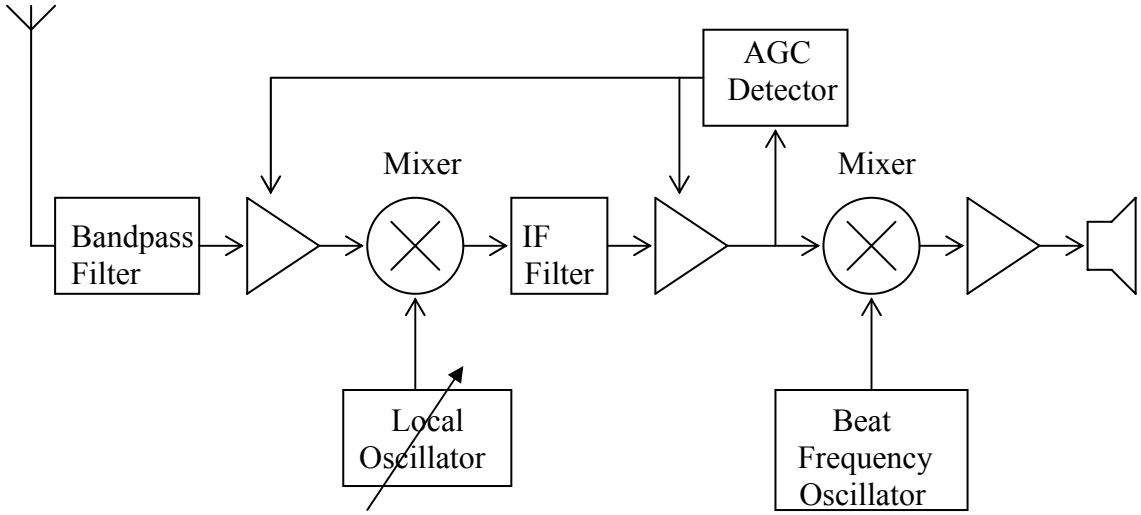
9 The output from a direct conversion receiver is the difference in frequency between:
a. The BFO and the incoming signal.
b. The BFO and the local oscillator.
c. The mixer and IF frequencies.
d. The incoming signal and the local oscillator.

10 A radio receiver that amplifies and filters the incoming signal at RF and then uses a diode detector to demodulate an AM signal would be called:
a. A superhet receiver.
b. A crystal set.
c. A tuned radio frequency receiver.
d. A direct-conversion receiver.

Chapter 23 - The Superheterodyne Receiver

The Single-Conversion Superhet

The superheterodyne receiver or "Superhet" as it is commonly known is the most widely used receiver design in amateur radio. It overcomes the lack of image rejection of the Direct Conversion receiver by converting the incoming RF signal to one or more *intermediate frequencies* before demodulating it. The block diagram of a typical *single-conversion superhet* (one with only a single intermediate frequency) is shown below.



A Single-Conversion Superhet Receiver

The RF signal from the antenna is first filtered by a band-pass filter. As in the Direct Conversion receiver this can be a fixed-tuned filter covering an entire amateur band, since the receiver does not rely on this filter (known as the *preselector*) for its selectivity. As we shall see, the main purpose of the preselector is to reject the image frequency. The signal is then amplified in an RF amplifier – once again, not too much amplification, to avoid overloading the mixer that follows (in some designs the RF amplifier may be omitted entirely).

In the first mixer, the RF signal is mixed with the signal from the tunable local oscillator. But instead of mixing it down to audio, this converts it to an *intermediate frequency* (IF). Common intermediate frequencies for single-conversion superhets are 455 kHz, 9 MHz and 10,7 MHz.

Suppose for example we want to receive a signal on 14,200 MHz again, and the intermediate frequency is 9 MHz. Then we could use a local oscillator frequency of either 5,200 MHz (because the difference between 5,200 MHz and 14,200 MHz gives the IF frequency of 9 MHz) or 23,200 MHz (because the difference between 14,200 MHz and 23,200 MHz is also the IF frequency of 9 MHz). For this example, we will assume that we chose a local oscillator frequency of 5,200 MHz, since this is within the range that can easily be generated by a VFO.

The resulting 9 MHz IF signal is then filtered by the IF filter, which is a very narrowband bandpass filter. Modern designs typically use crystal filters, so for this example we shall assume a crystal filter with a pass-band of 9,000 3 MHz (300 Hz above 9 MHz) to 9,003 0 MHz (3 kHz above 9 MHz). Signals within the pass-band will be passed with little

attenuation, while signals that fall outside the pass-band will be blocked. So what components of our original RF signal will fall within the filter pass-band? Well an RF signal at 14,200 3 MHz would be mixed down to 9,000 3 MHz by the 5,2 MHz local oscillator signal; and a signal at 14,203 0 MHz would be mixed down to 9,003 0 MHz. So the signals that originated at these frequencies – from 14,200 3 to 14,203 MHz – will make it through the IF filter. This corresponds to an USB signal at a frequency of 14,200 MHz.

What about signals on the "other side" of 14,200 MHz, from 14,197 0 to 14,199 7 MHz, i.e. the frequencies that would have caused an image in a Direct Conversion receiver? Well, they will be mixed down to between 8,997 0 MHz and 8,999 7 MHz, and will be rejected by the IF filter, so they do not cause a problem.

There is still an image, but in this case it is from 3,800 3 MHz to 3,803 0 MHz. A 3,800 3 MHz signal mixed with our 5,2 MHz local oscillator will generate an additive (sum) product at 9,000 3 MHz, and a 3,803 0 MHz signal will generate a mixing product at 9,003 0 MHz. So signals within the frequency range 3,800 3 MHz to 3,803 0 MHz when combined with the 5,2 MHz local oscillator signal will also generate products in the IF range from 9,000 3 to 9,003 0 MHz that will be passed by our IF filter. However this time the image is far away from the desired signal at 14,200 MHz, so it can easily be filtered out before the mixer, and this is the main purpose of the preselector. It must pass the desired frequencies, around 14,2 MHz, while rejecting the image frequencies, around 3,8 MHz. Fortunately because these frequencies are so far apart, it is fairly easy to get good "image rejection" from a simple bandpass filter made of inductors and capacitors.
To find the image frequency, just find the sum of, and difference between, twice the IF frequency and the desired receive frequency. So for the example above, with an IF of 9 MHz, twice the IF is 18 MHz. The sum of 18 MHz and the desired receive frequency of 14,2 MHz is 32,2 MHz. This is where the image would be if the design used a local oscillator with a frequency higher than the desired signal. The difference between twice the LO frequency, 18 MHz, and the desired receive frequency, 14,2 MHz, is 3,8 MHz, and this is where the image frequency will be with the local oscillator running at a lower frequency than the desired receive frequency, as it is in the example above.

Note that by varying the frequency of the local oscillator we can change what frequency RF signal will be mixed down to the 9 MHz IF. For example, a local oscillator frequency of 5,3 MHz would mix an RF signal of 14,300 MHz down to the 9 MHz IF, while our original reception frequency of 14,200 MHz would now be mixed down to 8,900 MHz and would be blocked by the IF filter. So can you tune a superhet receiver by varying the frequency of its local oscillator (the same as for a Direct Conversion receiver).

The circuitry after the IF filter is virtually identical to that of a Direct Conversion receiver. The IF signal is amplified, and then mixed with another locally generated oscillator signal – this time called the "Beat Frequency Oscillator" or BFO – to recover the audio signal, which is then amplified by an audio amplifier. Since the IF signal is at a fixed frequency – 9 MHz – the BFO does not have to be tunable so we can use a stable fixed-frequency 9 MHz crystal oscillator for the BFO.

The Automatic Gain Control (AGC) also works similarly to that of a direct conversion receiver, although in this case the AGC control voltage is derived from the intermediate frequency, rather than the audio frequency output. This gives us "IF-derived AGC" as opposed to the "audio-derived AGC" that we had in the direct-conversion design. IF-derived AGC is superior to audio-derived AGC as it is able to respond more rapidly to sudden changes in signal strength.

The same design can be used to receive CW signals as well. For example, to receive a CW signal with a frequency of 14,200 MHz, the local oscillator would be set to 5,199 4 MHz,

generating an IF signal at the difference between these frequencies, 9,000 6 MHz, which is within the pass-band of the crystal filter. After being amplified it will be mixed with the 9,000 MHz BFO signal in the product detector, generating an audio tone of 600 Hz.

So how about lower sideband signals? Well the simplest approach would be to have a second IF filter with a pass-band from 8,997 0 (3 kHz below 9 MHz) to 8,997 7 MHz (300 Hz below 9 MHz) that can be selected in place of the 9,000 3 to 9,003 0 MHz filter when we want to receive an LSB signal. Then when switching from USB to LSB all you have to do is switch filters, the local oscillator and BFO frequencies remain the same. Since crystal filters are quite expensive, an alternative approach is to use the same IF filter for LSB and USB reception, and just change the frequencies of the local oscillator and BFO. For example, to receive a LSB signal at 14,200 MHz using the 9,000 3 - 9,003 0 MHz IF filter we could set the local oscillator to 5,196 7 MHz and the BFO to 9,003 3 MHz. We leave it to the reader to fill in the details.

Since we can receive USB, LSB and CW signals using this design, how about AM signals? Well there are two options. The simplest is just to leave the receiver design exactly as it is, and receive AM signals as though they were single-sideband signals, ignoring the carrier and the other sideband, which will be filtered out by the IF filter. A better approach would be to provide another selectable IF filter, this time with a pass-band from 8,997 to 9,003 MHz to accommodate the 6 kHz bandwidth of an AM signal. The product detector would then be designed so that in the absence of any signal from the BFO, it would act as a half-wave rectifier and would detect AM by rectifying the IF signal (an “envelope detector”). This would give us the benefits of “proper” AM demodulation, notably accurate reproduction of the frequencies of the original audio signal even if the receiver is not perfectly tuned.

Multiple-Conversion Superhet Receivers

When choosing the IF frequency for a single-conversion superhet, there is a trade-off between image rejection and selectivity. It is easier to make highly selective filters at a low IF – say 455 kHz. However a low IF means that the image frequency is close to the desired frequency, making it difficult to effectively reject the image. Conversely, a high IF makes a large separation between the image frequency and the desired signal, making it easy to reject the image while passing the desired signal. However a high IF makes it harder to achieve the desired selectivity.

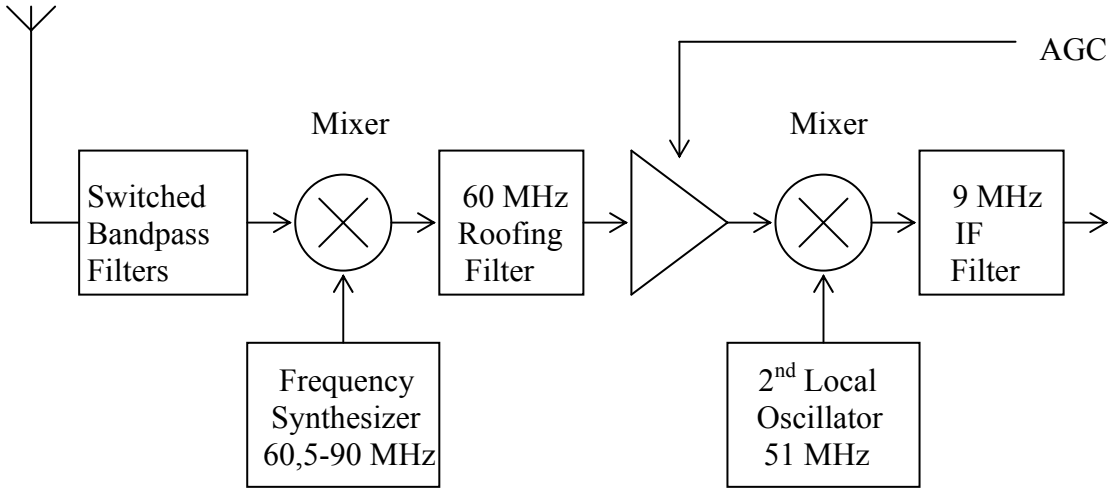
The classical solution to this dilemma has been to use a superhet design with *two* intermediate frequencies – a high *first IF* for good image rejection, followed by a low *second IF* for good selectivity. However modern crystal filters generally make this unnecessary in HF receivers, since very good selectivity is available from crystal filters at intermediate frequencies in the 9 MHz region, which is a high enough IF to attain good image rejection as well. Of course in VHF and UHF receivers, a higher first IF may be required to prevent unwanted image responses.

Despite this, the multiple-conversion superhet is still the most common approach for commercial HF receivers, but for a slightly different reason. Most commercial receivers and transceivers today offer “general coverage receive”, meaning that they can receive on any frequency in the MF and HF bands, typically from 500 kHz to 30 MHz. Unfortunately this gives them a problem with IF leak-through, which occurs when the first mixer is not exactly balanced, allowing some of the original RF signal to appear at the IF output. If the RF signal is at the same frequency as the IF, then it will be passed by the IF filter, causing the radio to respond to a frequency that it shouldn’t, a phenomenon known as a “spurious response”. This would not be a big problem for an amateur-bands-only receiver, because an IF frequency like 8,5 MHz could be chosen that is not close to any amateur band. Then the preselector, possibly assisted by a dedicated notch filter at the IF frequency, will be able to reject incoming RF

signals at the IF frequency, so there are no signals in the RF input that could "leak through" into the IF stages.

However the designer of a general-coverage receiver is not so fortunate. If the chosen IF frequency is anywhere in the receiver's frequency range, then it will be impossible to reject RF signals at the IF frequency, since these might include the frequency the receiver is tuned to! The solution is to choose an IF frequency that is either above or below the receiver's frequency range. However now the selectivity versus image rejection tradeoff comes back with a vengeance because a filter that is above the frequency range of a typical general coverage HF receiver – that is, above 30 MHz – will not have the necessary selectivity; while a filter at an IF that is below the receiver's coverage – say 455 kHz – will not allow adequate image rejection.

The usual solution is a multiple-conversion superhet where the first IF is *above* the receiver coverage range, allowing good image rejection and IF leak-through rejection, while the second IF is at a lower frequency where better selectivity can be obtained. This is known as an "up-conversion" design, since the incoming signal is first converted up to a higher frequency. The IF filter at the high first IF is often referred to as a *roofing filter* and is generally wide enough to permit signals of all modes through, up to 12 or 15 kHz in the case of a receiver that supports FM as well as other modes. Much narrower filters are provided for the different modes (e.g. a 6 kHz filter for AM and a 2,4 kHz filter for SSB) at the lower second IF. The block diagram below shows the "front end" (the circuitry from the antenna to the IF filter) of a typical general-coverage dual-conversion superhet.



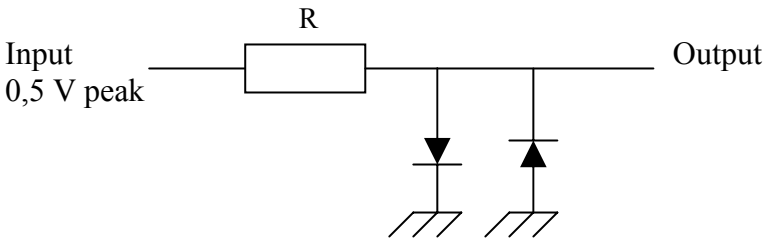
Front-End of a General Coverage Dual-Conversion Superhet

The design includes a bank of switched bandpass filters in the preselector, to allow coverage of the range 0,5 – 30 MHz with good image and IF leak-through rejection. The first local oscillator is a frequency synthesizer running from 60,5 to 90 MHz, which up-converts the RF signal to the first IF of 60 MHz. Here it is filtered by the roofing filter, which would typically have a bandwidth of 12 kHz or so. The purpose of the roofing filter is to reject signals which are close enough to the desired frequency to be passed by the preselector, but which might cause either inter-modulation distortion or an image response in the second mixer. The IF signal is then amplified and converted back down to the second IF frequency of 9 MHz. From here on the circuitry would be similar to the single-conversion design featured earlier.

Noise Limiters and Blankers

Many common sources of amplitude-modulated noise generate amplitude "spikes" of short duration but high amplitude, which extend over a wide range of frequencies. These may contain substantial energy due to their large amplitude, even though their duration is short. Such noise is generated both by natural sources, such as thunderstorms, and by man-made ones, like inadequately suppressed ignition systems. Interference from these noise sources can be reduced by noise limiters and blankers, which are available on almost all modern amateur transceivers.

A noise limiter is a very simple circuit that limits the maximum amplitude of the received signal.



Circuit Diagram of a Noise Limiter

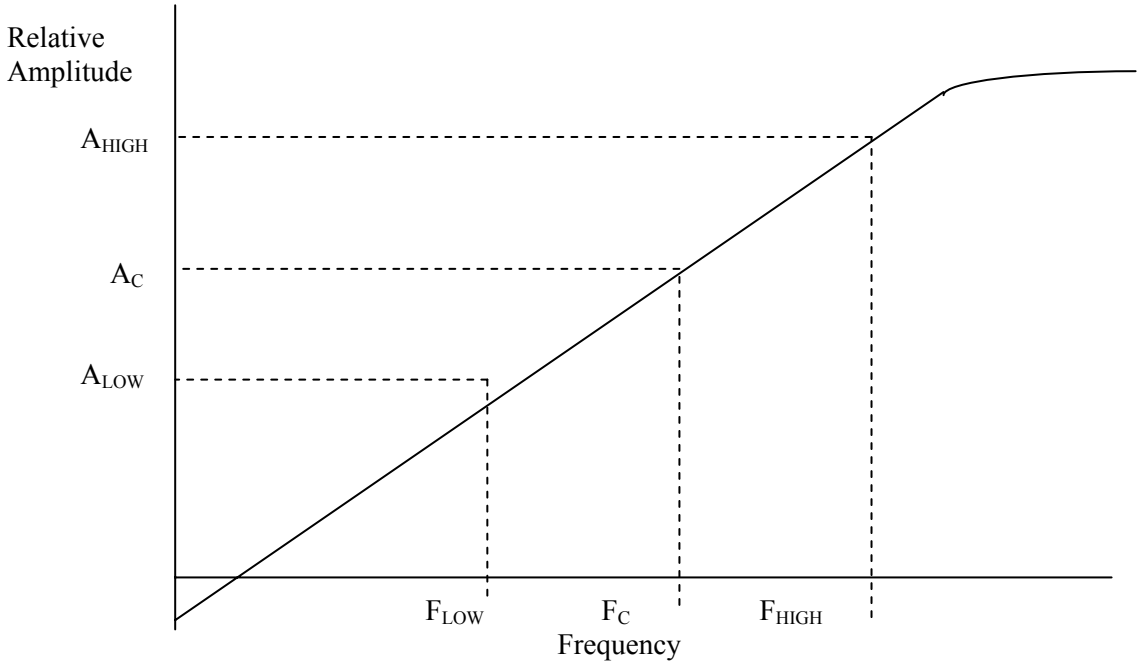
Assume the input signal has a maximum amplitude of 0,5 V peak under normal circumstances. This is less than the 0,6 V forward bias voltage of the diodes, so they do not conduct, and the input signal will be passed to the output unchanged. Then suppose a noise pulse generates a signal amplitude of 5 V. As soon as the amplitude exceeds 0,6 V, the diodes will conduct, effectively limiting the maximum output to 0,6 V peak and substantially reducing the energy of the noise signal.

The noise blanker is a more sophisticated variation on this idea. It detects the large amplitude of the incoming noise signal, and then immediately mutes (turns off) the audio output of the receiver completely for a predetermined time, typically a few milliseconds. Although this blocks the desired signal as well as the noise, this usually goes unnoticed by the listener as the human ear is quite insensitive to very short gaps in sounds, and the resulting signal degradation is much less than would have been caused by the high amplitude noise spike.

Frequency Modulation (FM) Reception

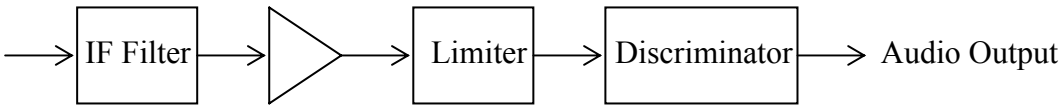
The basic superhet design can also be used to receive frequency modulated (FM) signals. However in this case, the product detector is replaced by a *Foster-Seeley discriminator* or a *ratio detector*. These are circuits that convert frequency variations into a varying output voltage, so recovering the modulation from an FM signal.

The discriminator works by positioning the FM signal on the slope of a selective filter, so that variations in the frequency of the FM signal will result in variations in its amplitude. This converts the frequency modulation into a combined amplitude and frequency modulation, and a diode detector is used to recover the modulation from the AM component.



The graph shows how the slope of a high-pass filter could be used to convert frequency modulation into amplitude modulation. As the signal frequency increases from F_C, the centre frequency, to $F_{HIGH,}$ the amplitude of the output increases from A_C to A_{HIGH}. If the frequency decreases from F_C to F_{LOW}, then the amplitude of the output will also decrease, from A_C to A_{LOW}.

Because the discriminator is also sensitive to changes in the amplitude of the incoming signal, it should be preceded by a *limiter*. This is a circuit that limits the amplitude of the signal, so that amplitude variations are not passed on to the discriminator or ratio detector that follows. The limiter circuit is identical to the noise limiter discussed earlier, except that in an FM receiver the circuit would be driven at a much higher input level, causing the diodes to conduct and clamp the output signal to 0,6 V peak. In this way the output of the limiter will always be at the same level (0,6 V peak), irrespective of the amplitude of the input signal. The block diagram below shows the final IF stage of a typical FM receiver



Final IF Stage of an FM Receiver

When the received signal is very weak the limiter is ineffective and the discriminator will respond to amplitude variations, which cause hiss in the audio. As the signal gets stronger and the limiter takes effect, the hiss decreases, a process called “quieting”. In order to prevent the hiss from bothering the listener when there is no received signal, most FM receivers incorporate a squelch feature, which mutes (turns off) the audio output when the received signal is below a minimum level known as the squelch threshold. The squelch threshold may be fixed or it may be adjustable using a squelch control.

Summary

The superhet receiver converts the incoming RF signal to one or more *intermediate frequencies* before demodulating it. Superhet receivers have an *image frequency* that when mixed with the local oscillator will also generate the same IF as the desired receive signal. The image frequency will be either the sum of, or the difference between, twice the IF frequency and the desired receive frequency. The role of the preselector is to reject incoming RF signals at the image frequency, preventing them from causing a spurious (unwanted) response in the receiver. The choice of intermediate frequency is a tradeoff between selectivity (better at low frequencies) and image rejection (better with a higher frequency IF). If a single IF cannot give adequate selectivity and image rejection, then a dual conversion design may be employed, with a higher first IF to give good image rejection, and a lower second IF to give good selectivity.

Noise limiters limit the amplitude of pulse noise, reducing the effect on the receiver. Noise blankers mute the audio output for a short time (a few milliseconds) when the higher amplitude associated with pulse noise is detected.

FM signals are detected using a *Foster-Seeley discriminator* or *ratio detector*. The discriminator should be preceded by a limiter to prevent it from being affected by variations in the amplitude of the signal. Weak FM signals have a characteristic hiss on them, and as the signal strength increases and the limiter becomes effective the hiss goes away, a process known as *quieting*. Most FM receivers incorporate a *squelch* function, which mutes the audio output when there is no received signal to avoid the annoying hiss.

Revision Questions

1 In an FM receiver the effect of sufficient signal arriving to start the limiter operating, thus reducing background noise, is known as:
a. Damping.
b. Squelch.
c. De-emphasis.
d. Quieting.

2 The selectivity of a receiver is mostly controlled by:
a. Gain of IF and RF stages.
b. Bandwidth of RF and IF stages.
c. Sensitivity of RF and IF stages.
d. Stability of RF and IF stages.

3 In a superheterodyne receiver intended for AM reception, what stage combines the received radio frequencies with energy from a local oscillator to produce an output at the receiver's intermediate frequency?
a. The mixer.
b. The detector.
c. The RF amplifier.
d. The AF amplifier.

4 In superheterodyne receivers the setting of the first IF is governed by two general principles:
a. High IF gives good image rejection but low IF gives better selectivity.
b. High IF gives good image rejection and good selectivity.
c. Low IF gives good image rejection and high IF gives good selectivity.
d. Low IF gives good image rejection and good selectivity.

5 The function of an IF amplifier in a superheterodyne receiver is to:
a. Improve its sensitivity.
b. Improve its selectivity.
c. Buffer the mixer output.
d. Amplify the loudspeaker output.

6 How can the selectivity of an IF amplifier be improved?
a. Varying the supply voltage.
b. Varying its resistance.
c. By use of a bandpass filter.
d. By use of a Low-pass filter

7 The detection of a Single Side-band signal in a receiver requires a:
a. Carrier Insertion Oscillator.
b. Special Aerial.
c. An SSB amplifier.
d. A special transformer.

8 A superheterodyne receiver is operating with its local oscillator on the high side of the incoming signal. If its IF is 450 kHz and it is receiving an input signal of 14 100 kHz an image will be produced:
a. At this tuning point if a strong signal on 15 000 kHz is present.
b. Further up the tuning band if a strong signal on 15 000 kHz is present.
c. At this point if a strong signal on 13 650 kHz is present.
d. Further up the tuning band if a strong signal on 13 650 kHz is present.

9 In a single conversation superheterodyne receiver with an IF of 450 kHz a signal is received first at 12 000 kHz and then again at 12 900 kHz. These two received signals are called:
a. Cross-modulation products.
b. Band spread products.
c. Image signals.
d. De-emphasis signals.

10 The process by which a receivers' local oscillator and mixer resonant circuits maintain a constant IF frequency separation is known as:
a. Tracking.
b. Isolation.
c. Shielding.
d. Attenuation.

11 The preferred circuit for resolving an SSB signal is:
a. A product detector.
b. A fullwave rectifier.
c. A Colpitts oscillator.
d. A crystal oscillator.

12 The product detector is used to:
a. Detect square waves.
b. Balance out noise signals.
c. Deduce unwanted feedback.
d. Resolve SSB and CW modulation.

13 What is the purpose of the detector in a receiver?
a. To amplify the incoming signal.
b. To operate the squelch circuit.
c. To operate the on/off light.
d. To reproduce the modulating signal.

14 Electrical Interference or reception can best be limited by means of a:
a. Squelch circuit.
b. Noise Limiter.
c. Isolation transformer.
d. Decoupled loudspeaker.

15 The receive facility that switches the audio circuit off in the absence of a satisfactory signal strength is:
a. A Noise limiter.
b. A Squelch circuit.
c. A VOX circuit.
d. An AF gain control.

16 In order to avoid image reception on VHF receivers they normally have:
a. Low IF frequencies.
b. Crystal controlled local oscillators.
c. A stable BFO.
d. High IF frequencies.

17 A Dual conversion receiver contains:
a. Two IF amplifiers of different frequencies.
b. Two RF pre-amplifiers.
c. Stereo audio circuits.
d. Two antenna connections.

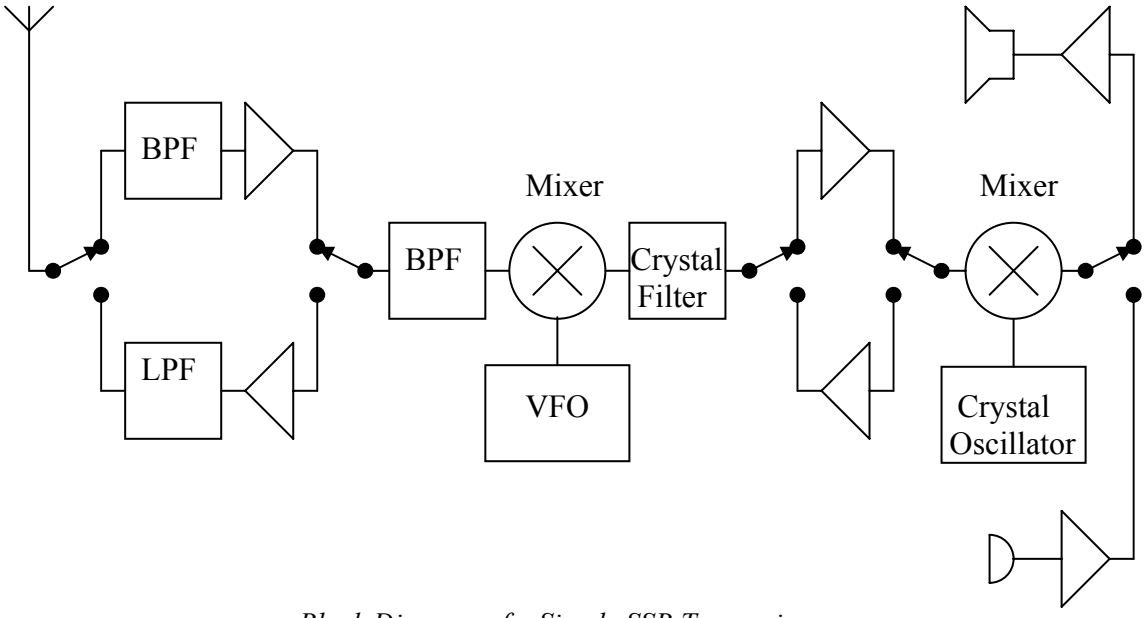
Chapter 24 - Transceivers and Transverters

Although in the early days of amateur radio transmitters and receivers were usually separate, in modern practice these are usually combined in a single piece of equipment, the *transceiver*. Transceivers have several advantages over separate transmitters and receivers:

- It is easier to make the transmitter and receiver tune together, so the user does not have to separately tune the transmitter and receiver to the same frequency.
- Much of the circuitry including the oscillators or synthesizers, filters, antenna switching, microprocessor and display can be shared between the transmitter and receiver, making transceivers less expensive than a separate transmitter and receiver.
- Installation is simpler, with less space and fewer cables required.

For these reasons, almost every modern amateur transmitter also includes receive capability, at least on the bands that it can transmit on. Many transceivers also offer "general coverage" receive, being able to receive signals outside the amateur bands.

In order to maximize the sharing of components between the transmitter and receiver section of a transceiver, they will typically use the same frequency conversion scheme but in reverse. For example, if the receiver is a double-conversion superhet with IFs at 60 MHz and 9 MHz, then the transmitter will probably generate SSB using the filter method at 9 MHz (Allowing reuse of the 9 MHz crystal filter), and then mix it up to 60 MHz using the receiver's beat frequency oscillator, and then mix it back down from 60 MHz to the actual transmit frequency (allowing the same synthesizer to be used for both transmit and receive). Circuit reuse is further enhanced since popular balanced mixers (such as the passive diode mixers) are essentially reversible – a signal injected at the RF port will mix with the local oscillator to generate a signal at the IF port, while a signal injected at the IF port will mix with the local oscillator to generate a signal at the RF port. Many filters will also work equally well in either direction. The diagram below shows a simple SSB transceiver that reuses several of its functional blocks. "LPF" stands for "Low-Pass Filter", "BPF" for "Band-pass Filter" and "VFO" for "Variable Frequency Oscillator".



Block Diagram of a Simple SSB Transceiver

There are two different signal paths through the transceiver, depending on the position of the transmit/receive switches (these would probably be implemented using electronic switching or relays). With the switches in the position shown, the transceiver is in receiving mode. The signal from the antenna is filtered by a band-pass filter, amplified in the IF amplifier, fed through another band-pass filter and then mixed down to IF by the signal from the VFO. This passes through a narrow crystal filter which removes all frequencies other then the desired one, is amplified in the RF amplifier and finally demodulated in the product detector, using the signal from the crystal oscillator that serves as a beat frequency oscillator (BFO) for the receiver.

On transmit (with the switches all switched the other way) the signal from the microphone is amplified by the preamplifier, and then fed into the demodulator, which this time serves as a balanced modulator. The output of the balanced modulator is amplified, filtered using the crystal filter to remove the unwanted sideband, and mixed to the final output frequency using the signal from the VFO. The output is passed through a band-pass filter to remove the unwanted mixing product, then amplified by the RF power amplifier and finally filtered to remove any harmonics. This design reuses the mixer, product detector, crystal filter, VFO, crystal oscillator and one of the band-pass filters.

Most amateur transceivers are designed to operate into an unbalanced antenna with an impedance of 50 Ω.

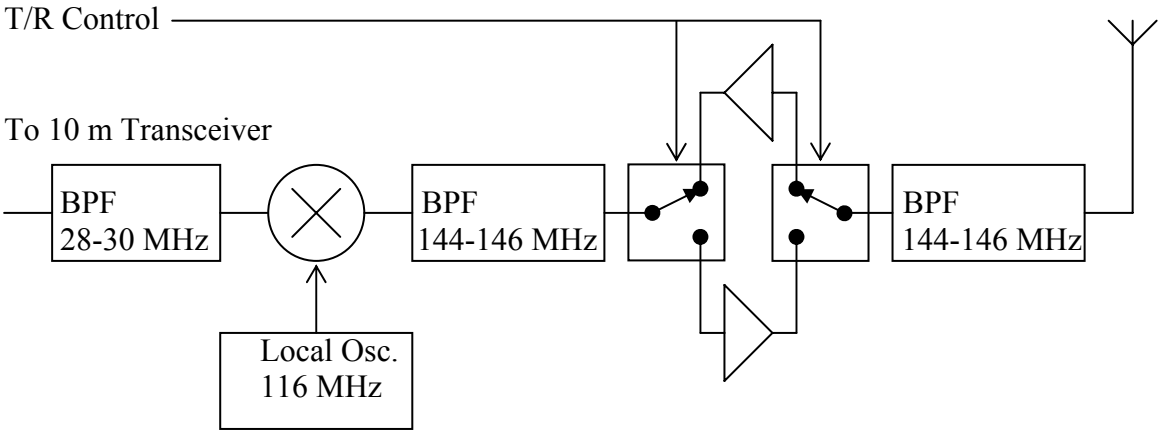
The Transverter

A *transverter* is a device that allows a transceiver to transmit and receive on a frequency band other than the ones it was originally designed for. It does this by incorporating a local oscillator and mixer that can convert the output of the transceiver to the new frequency band, and convert signals received on this band to a frequency that the transceiver can receive on.

For example, a 2 m transverter might convert frequencies in the range 28 – 30 MHz (around the 10 m amateur band) to the frequency range 144 – 146 MHz (the 2 m band). It could do this by mixing the signals with a 116 MHz local oscillator signal. On transmit, the sum of the local oscillator and an input signal in the range from 28-30 MHz would give an output in the range 144-146 MHz, while on receive the difference between an input signal in the range 144-146 MHz and the 116 MHz local oscillator signal would give an output in the range 28-30 MHz.

Most transverters are designed to operate with low powers on transmit, generally around 0 dBm (1 mW), and include a power amplifier to amplify the signal. Some transceivers have a special connector for transverters, which provides a low-level RF signal to drive them. Otherwise an attenuator should be used on transmit to decrease the transceiver's output to a level that the transverter can safely handle.

Like transceivers, transverters usually use components like the local oscillator, mixer and filters for both transmit and receive. Their transmit/receive switching is usually controlled by the transceiver, using its T/R control output. A block diagram of a typical 2 m transverter is shown below. It includes both a power amplifier for transmit, and a receive preamplifier to amplify weak signals and compensate for losses in the mixer.



Block Diagram of a 2 m Transverter

On transmit, the low-level signal from the transceiver is filtered by the 28-30 MHz band-pass filter, and then mixed with a 116 MHz local oscillator. This generates a "sum" product in the range 144-146 MHz, and a "difference" product in the range 90-88 MHz. The band-pass filter that follows the mixer rejects the difference product. The 144-146 MHz product is amplified by a power amplifier, and passed through a final band-pass filter to remove any harmonics.

On receive, the signal from the antenna is filtered by the 146-148 MHz band-pass filter, amplified by a low-noise preamplifier, and filtered again to remove any image signals in the 88-90 MHz range. It is then mixed with the 116 MHz local oscillator, generating a "sum" product at 262-264 MHz, and a "difference" product in the range 28-30 MHz. The 28-30 MHz band-pass filter rejects the unwanted "sum" product, leaving only the 28-30 MHz signal that is fed to the transceiver.

Note that transverters typically translate frequencies from anywhere in a whole band, to the corresponding place in a different band, so the transverter does not need a tunable local oscillator – the tuning will be done in the transceiver that is connected to it. Some transceivers allow the frequency display to be offset when using a transverter so for example, the transceiver display could read 144-146 MHz while the transceiver was actually being tuned from 28-30 MHz, correctly indicating the output frequency of the transverter.

Summary

A transceiver consists of a transmitter and receiver combined into one. They are widely used because of operator convenience and the lower cost that can be achieved by using the same components for both transmit and receive functions.

Transverters convert a transceiver to transmit and receive in a new frequency band. They work by mixing the output of the transceiver or the input from the antenna with a fixed frequency local oscillator, translating the frequency to the new band. Transverters generally require an input signal power of about 1 mW when transmitting, and care should be taken not to overdrive them.

Chapter 25 - Antennas

Antennas convert electrical energy – which requires conductors to carry it – into electromagnetic energy, which is able to radiate through space.

An electrical current flowing in a conductor generates a magnetic field in the space around the conductor. This is the principle behind electromagnets. If the current flowing in the conductor varies with time, then the magnetic field around the conductor will also vary with time. However according to the principle of induction a varying magnetic field will give rise to an electric field; and conversely, a varying electric field will give rise to a magnetic field. So by varying the current in a conductor, we can create a varying magnetic field, which will in turn create a varying electric field, which will create a varying magnetic field, and so on. The resulting interrelated varying electric and magnetic fields are called electromagnetic waves, and can travel long distances. At the frequencies that we are interested in, these electromagnetic waves are radio waves, although heat, light, and x-rays are also examples of electromagnetic waves of higher frequencies.

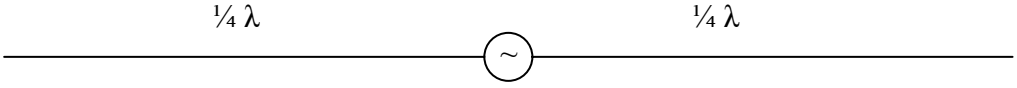
In electromagnetic waves, the electric and magnetic fields are perpendicular (at right angles) to each other, and both are perpendicular to the direction of motion of the wave. So if the electric field is horizontal with respect to the surface of the earth, the magnetic field will be vertical; while if the electric field is vertical, then the magnetic field will be horizontal. We refer to the *polarization* of electromagnetic waves according to the orientation of the electric field. So if the electric field is horizontally oriented, then the wave is horizontally polarized; while if the electric field is vertically oriented, it is vertically polarized. In physics, the electric field is referred to as the E-field, and the magnetic field as the H-field.

The orientation of the electric field (and hence the polarization of the radio waves) corresponds to the orientation of the conductor carrying the current that generated the radio waves, so an antenna consisting of horizontal conductors will generate horizontally polarized radio waves, while an antenna consisting of vertical conductors will generate vertically polarized radio waves.

Polarization is important because a vertically polarized antenna will not respond to horizontally polarized radio waves, and vice-versa. So for line of sight communications it is important that the polarization of the transmitting antenna should be the same as that of the receiving antenna.

The Half-Wave Dipole

The half-wave dipole is a simple antenna that consists of a ½ wavelength of wire that is fed in the centre by a radio-frequency voltage source.



A Half Wave Dipole

Assume that an alternating voltage is applied at the dipole's resonant frequency – that is, the frequency for which each side of the dipole is exactly ¼ wavelength.

Consider an instant in time when one side of the voltage source is positive and the other side is negative. The effect will be to pull some electrons out of the side of the dipole that is attached to the positive terminal of the voltage source, and push some electrons into the side

that is attached to the negative terminal of the voltage source. Since electrons repel each other, as you force electrons into the wire at the negative terminal of the voltage source, they will repel the electrons that are already in the wire, pushing them towards the end of the wire. As each electron moves up a bit, it will repel its neighbour, forcing it to move up too, and so on causing a wave to travel down the wire. This wave will travel at the speed of light until it reaches the end of the wire, where the electrons can no longer bunch up any more (because they have nowhere to move). At this point, the wave will reflect from the end of the wire and head back towards the feed-point.

The effect is similar to what you would observe if you set up a line of pool balls with one end against the cushion, and then knocked the one furthest from the cushion into its neighbour. Each ball hits the next one, and so on down the line, until the last ball, which is up against the cushion so it cannot move. The "pool-ball wave" is then reflected from the cushion and travels back in the opposite direction. Of course the pool-ball wave does not travel at the speed of light!

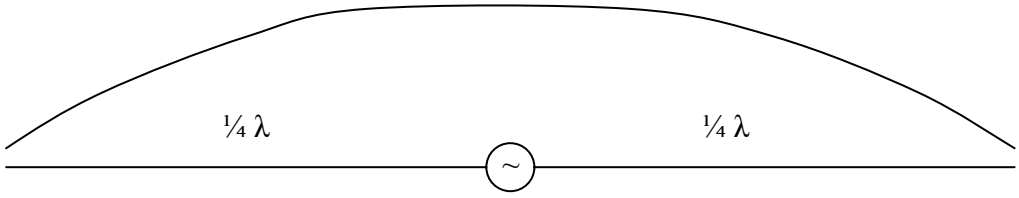
Similarly in the dipole, applying an instantaneous potential difference at the feed-point (the place where the voltage source is connected tot he antenna) causes waves to travel from the feed-point to both ends of the wire, where they are reflected and head back towards the feed-point. On one side – the side where the negative potential is applied – the wave consists of an increase in the electron density, "compressed electrons" (like the pool-ball analogy). On the other side – the side where the positive potential is applied – the wave consists of a reduced electron density as electrons have been attracted out of the wire by the positive potential.

These waves both have to travel ¼ wavelength to the end of the wire, and another ¼ wavelength in the opposite direction before they get back to the feed-point, a total distance of ½ wavelength. So one half-cycle later, the waves will return to the feed-point, heading in the opposite direction. But because it is half a cycle later, the voltage source will have the opposite polarity, so it will now be pushing the electrons in the opposite direction. But this is the same direction that the reflected wave is traveling in, so the reversed polarity of the voltage source will reinforce the reflected wave. Another half cycle later the waves will have reflected off the ends of the wire again, and the voltage source will again have reversed polarity, so once again the voltage source will reinforce the waves of increased and reduced electron density that are coursing up and down the wire like water sloshing about in a bath.

Because the applied voltage is always reinforcing the waves, a fairly small voltage can (over a few cycles) cause a large movement of electrons, in other words a large current (since the electrons are charge carriers that are flowing backwards and forwards, and so constitute an alternating current flowing in the antenna). Remembering that from ohm's law, $R = V/I$ we see that the feed-point resistance will be low, since a small V causes a large I. In actual fact, the feed-point impedance of a half-wave dipole is about 72 Ω.

The fact that the resistance is not zero means that the antenna is dissipating power. Where is this power going? It is being radiated as radio waves from the antenna. This apparent resistance of the antenna caused by energy being radiated from it is called the *radiation resistance* of the antenna.

You will also note that in the centre of the antenna, the electrons are quite free to move, so a large current will flow. However the nearer you get to the ends of the antenna, the less free the electrons are to move, up to the points right at the ends, where the electrons can hardly move at all. This means that there will be a larger current flowing in the centre of the antenna than at the ends. If you superimpose a graph of the amplitude of the current flowing on top of a diagram of the antenna, you get something like this:



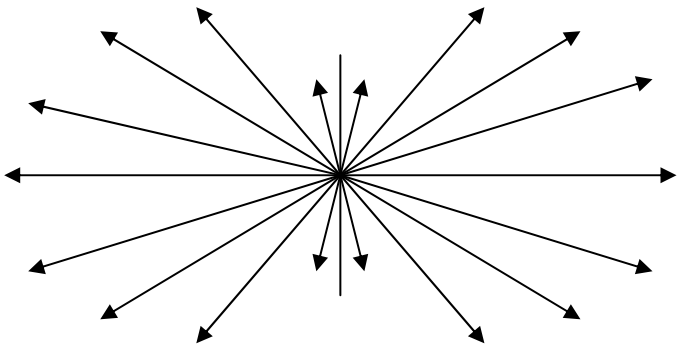
The Current Distribution in a Half-Wave Dipole

As you can see, the current is strongest in the centre of the dipole and tapers off towards the ends. Because it is the current flowing in the antenna that is primarily responsible for the emission of radio waves, they can be visualized as being emitted from the point of highest current – called a current *loop* – at the centre of the antenna. This has some practical implications – for example, the ends of a dipole can be bent without affecting its properties as an antenna much, since the ends are relatively unimportant as far as radiation is concerned.

Although the current in a dipole decreases towards the ends, the opposite happens with the voltage – it is greatest at either end of the dipole. In general, points of high current (current *loops*) are points of low voltage (voltage *nodes*), and points of low current (current *nodes*) are points of high voltage (voltage *loops*). So beware the open ends of antennas, where no current is flowing, as they invariably carry high voltages!

This current distribution – and the corresponding voltage distribution – are called "standing waves" since they have a wave-like shape (roughly like a sine wave) but the points of highest and lowest current and voltages do not move – they are standing still. Another way to think of the standing waves is that they are caused by the interaction between the waves from the feed-point moving towards the ends of the antennas, and the reflected waves moving back towards the feed-point.

Our next task is to calculate in what directions the dipole will radiate; this is called the *radiation pattern* of the antenna. This is quite easy for an antenna that has only a single point of maximum current, like the dipole. When an alternating current flows in a wire, it radiates most strongly at right angles to the wire, and less strongly the further you move from the right-angled direction. This is depicted below, where the length of each arrow represents the strength of the radiation coming from the wire in every direction.

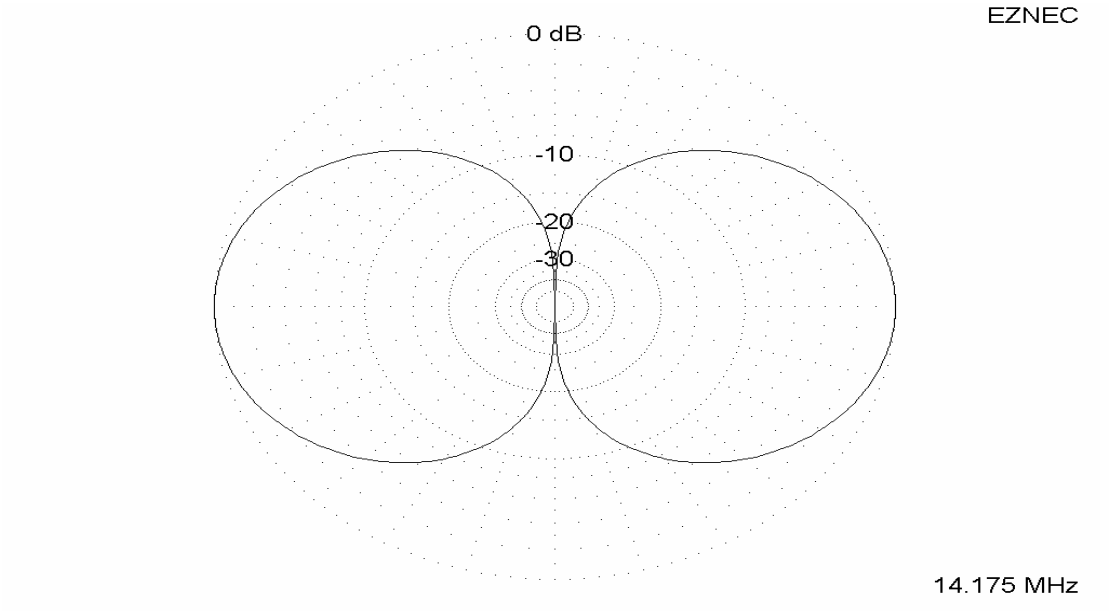


Radiation Pattern of a Dipole

In this case the dipole is the vertical line in the centre of the diagram. Note how the strongest radiation is perpendicular (at right angles) to the wire, and the strength of the radiation

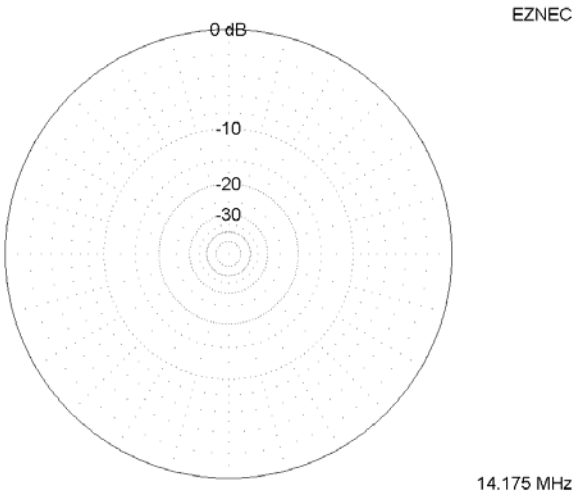
decreases as the angle gets further from perpendicular. There is no radiation at all along the axis of the wire (i.e. vertically up or down the page in this diagram).

Actually we don't usually draw radiation patterns with arrows like this. Instead we just draw a line that would join the tips of all the arrowheads. In other words, the distance from the centre of the diagram to the line indicates the strength of radiation in that direction. The following diagram is the radiation pattern of a dipole, drawn in the conventional way.



Radiation Pattern of a Dipole in Free Space

Although the dipole is not shown, it is oriented the same as in the previous diagram – vertically on the page in the centre of the plot. The plot lines indicate the relative strength of the field in each direction: the further the distance from the centre of the diagram to the line, the stronger the radiation in that direction. Note the nulls (points of minimum radiation) off the ends of the dipole (vertically up and down in the plot). Of course, the dipole will radiate equally in all directions perpendicular to it, not just in the two directions (left and right) shown in the diagram. So if you think of it in three dimensions, it would look like a doughnut shape with the wire in the middle. We can also look at the dipole end-on. Seen end on, the pattern is like this:

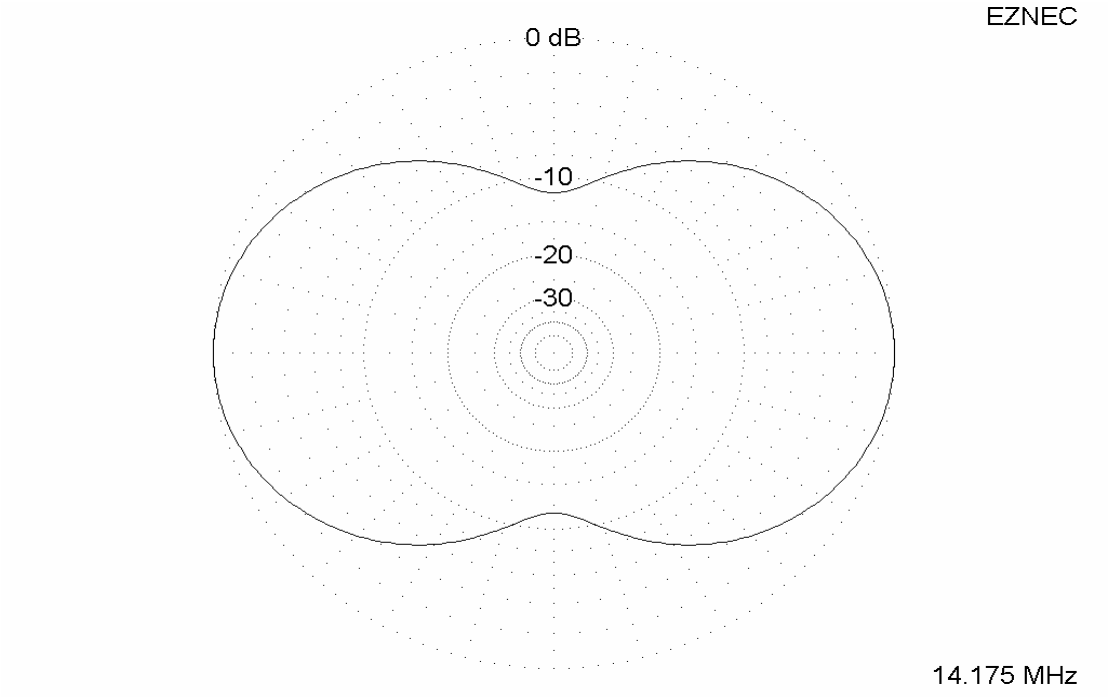


Radiation pattern of a Dipole in free space viewed end-on

Here the dot in the middle represents the wire seen end-on, and the circular radiation pattern indicates that it radiates equally in all directions.

These diagrams show the radiation pattern of a dipole in *free space*, i.e. far away from the ground. The ground reflects radio waves, so in order to understand the radiation pattern of antennas mounted over ground (like all normal amateur antennas), we also need to take into account the effect of these reflections. In general, the waves reaching a distant point will come from two sources – a direct wave from the antenna, and a wave that has been reflected by the ground. Depending on the difference between the distances traveled by these waves, they may reinforce each other, or they may cancel each other out.

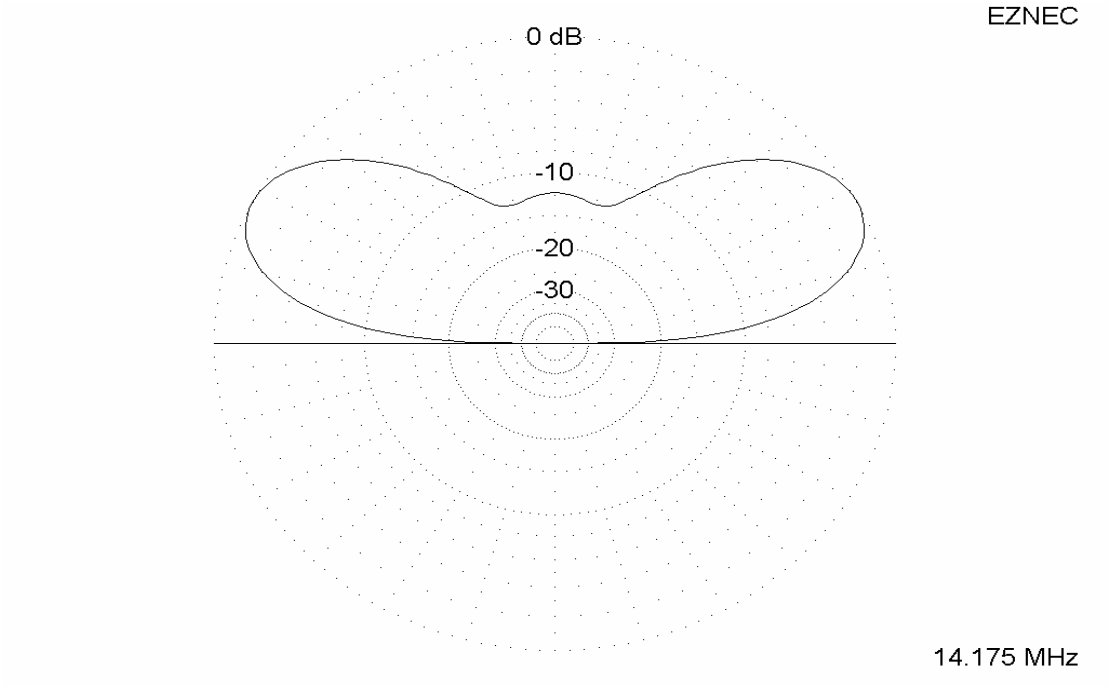
Assume that we orient the dipole horizontally and mount it some distance above ground. Then the pattern viewed from above, would look pretty much the same as the free-space pattern.



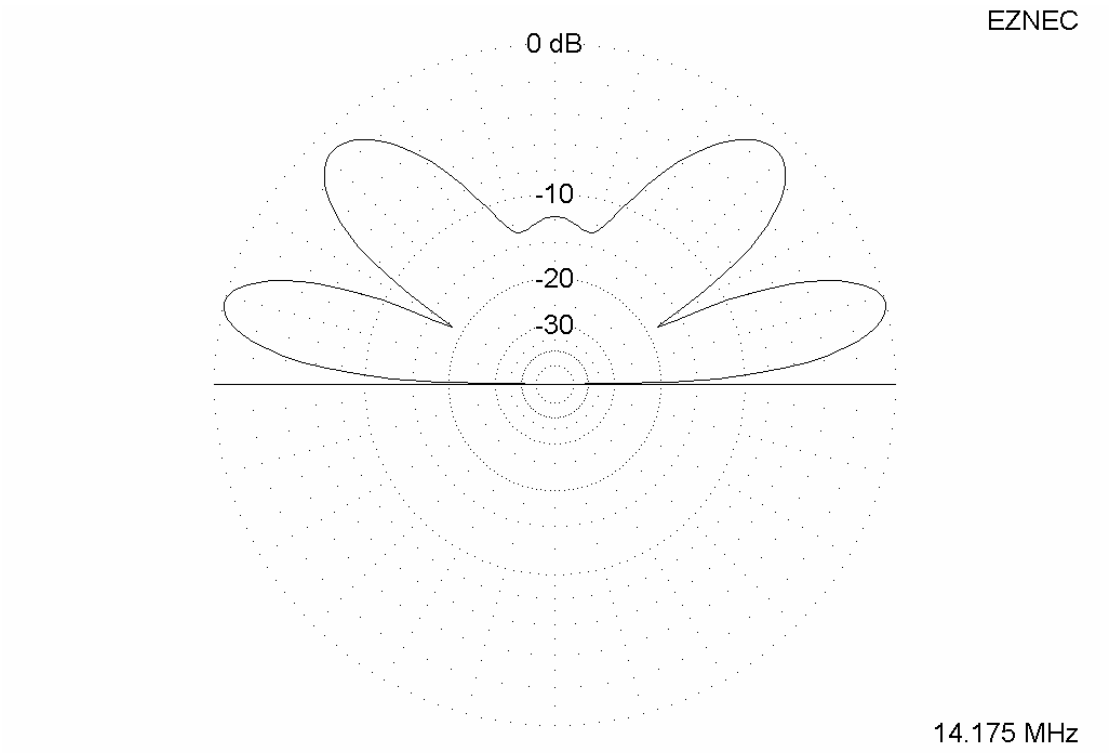
Azimuthal radiation pattern of a horizontal half-wave dipole near ground.

Once again the dipole is oriented vertically on the page, in the middle of the diagram, although this time we are looking down on it from above. This is what we call an "azimuth pattern". The main effect of the ground reflections has been to "fill in" the nulls (directions of zero radiation) off the ends of the dipole, although there is still much less radiation from the ends of the wire than perpendicular to it – approximately 13 dB less according to the scale on the diagram. For this reason, we talk about the horizontal dipole as being a *bi-directional* antenna, favouring two directions (left and right in the diagram above) at the expense of the others.

The nearby earth has a much greater effect on the vertical pattern of the antenna. Instead of radiating equally in all directions, as it would with no ground present, we get the following pattern:



The ground reflections have cancelled out most of the low-angle and high-angle radiation, leaving one "lobe" on each side, with an angle of maximum radiation of about 27º in this instance. The exact pattern depends on the height of the antenna above ground. In this case, it is ½ wavelength above ground. If we raise it to 1 wavelength, the elevation pattern now looks as follows:



There are now two lobes in each direction, one at a fairly low angle (14º) and one at a higher angle (47º). In general, the higher we raise the dipole above ground, the more lobes we get, and the lower the elevation angle of the lowest-angle lobe.

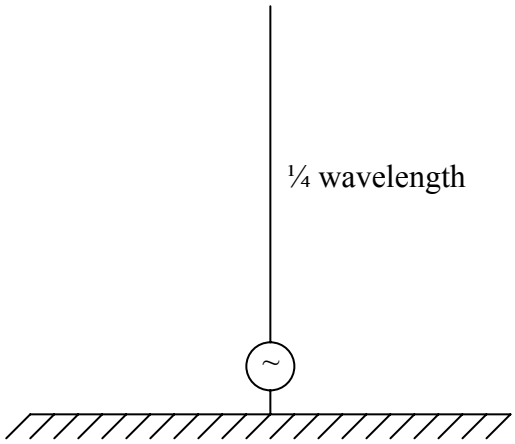
Low-angle radiation is very desirable for making long-distance contacts, so if making long distance contacts is a priority then the higher your antenna, the better, at least for horizontally-polarized antennas like this dipole.

The dipole is normally used as a horizontal antenna, but it can also be constructed vertically if desired, especially in the VHF and UHF bands where the wavelength is shorter and the height required more reasonable. A vertical dipole has the same radiation pattern as the quarter-wave vertical discussed below.

A horizontal half-wave dipole is sometimes called a "Hertz" antenna, since this type of antenna was first used by Gustav Hertz, the German physicist who also gave his name to the unit of frequency.

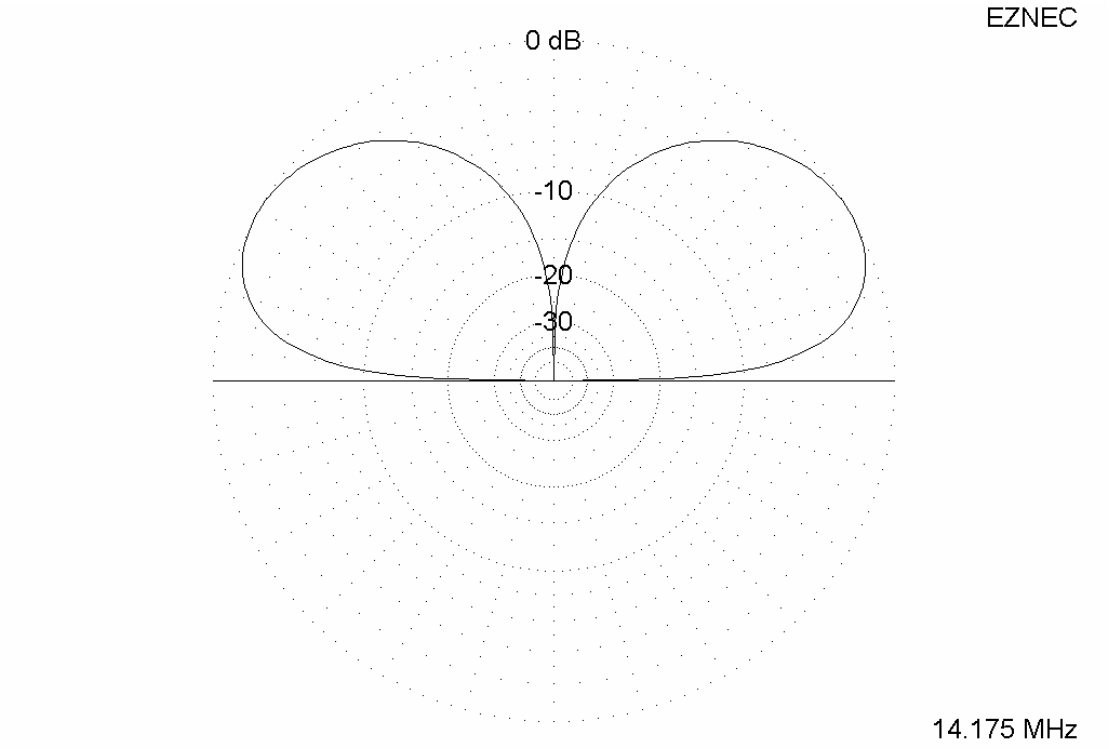
The Quarter-Wavelength Vertical

Another popular antenna is the quarter-wavelength vertical. It consists of a ¼ wavelength of wire mounted vertically and driven at its base, with other side of the driving voltage connected to ground. It works like "one half of a dipole", with the return path for the current being through the ground.



A Quarter-Wavelength Vertical

There's not much point in talking about the radiation pattern of a ¼ wavelength vertical in free space, since it needs the ground as one of its connections! The elevation plot of the quarter-wave vertical over ground is shown below.



Elevation pattern of a quarter-wavelength vertical

The azimuth pattern (the pattern when viewed from above) would just be a circle, since the antenna radiates equally in all directions when viewed from above. Antennas that radiate equally in all (horizontal) directions are called *omni-directional* antennas.

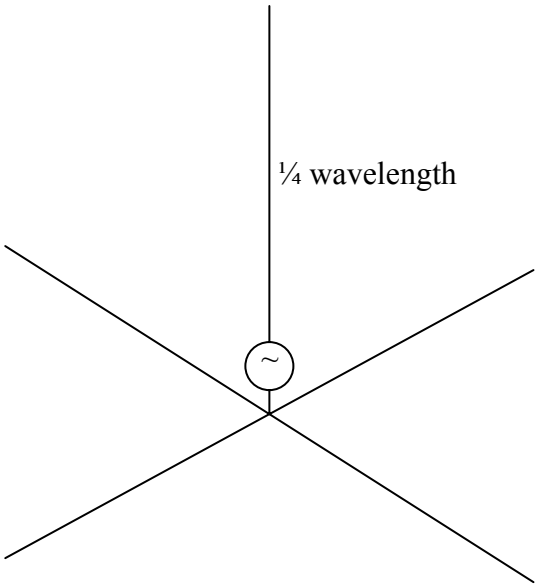
Although simple in theory, the quarter-wave vertical has a difficult practical problem to overcome: it is difficult to create low-impedance ground connections at radio frequencies. A ground rod of the kind used for household mains grounds will generally not present a very low impedance at radio frequencies, and the resistance of the resulting ground connection will cause power to be dissipated as heat in the ground rather than being radiated from the antenna. Although a quarter-wave vertical will work with just a ground-rod, you could easily find 75% or more of the power applied to the antenna being wasted heating the ground instead of being radiated.

The radiation resistance of a ¼ wavelength vertical is about 36 Ω. However the effective ground resistance will also contribute to the impedance seen at the feed-point, which will usually be quite a bit higher than this.

A ¼ wave vertical is sometimes called a "Marconi" antenna, since this type of antenna was first used by Guglielmo Marconi, the inventor of wireless telegraphy and pioneer of transatlantic radio communication.

The "Ground Plane" Antenna

One solution to the problem of creating a low-impedance RF ground connection is to raise the vertical antenna 1/8 wavelength or more above the ground, and feed it against 3 or 4 quarter-wavelength radials, rather than against ground. The radials effectively act as the "missing side" of the dipole, but because they run in opposite directions, the radiation from them cancels, leaving only the radiation from the vertical wire, which is called the *radiator*.



A "Ground Plane" Antenna

The radials in a ground-plane antenna may be laid out flat (as shown in the diagram) or they can droop downwards. With flat radials, the radiation resistance will be between 20 Ω and 26 Ω depending how high above ground the antenna is mounted. Drooping the radials will increase the feed-point impedance, so with the right angle of droop (about 45º) the impedance can be raised to 50 Ω, which is a good match for the coax cables usually used to feed such antennas. The radiation pattern of a "ground-plane" antenna is identical to that of a quarter-wave vertical.

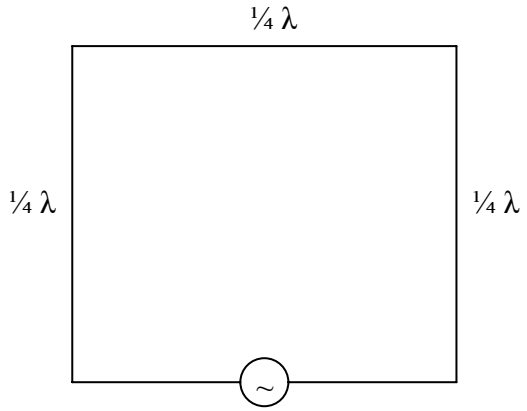
Short Antennas

An antenna that is shorter than the length required for resonance will have some capacitive reactance. For example, a vertical antenna that is somewhat less than ¼ wavelength will have some capacitive reactance. This can be cancelled out by an equivalent amount of inductive reactance in series with the antenna. This may be provided by an inductor known as a *loading coil*, which may be placed at the base of the antenna (*base loading*) or in the centre of the antenna (*centre loading*). Centre loading is more efficient electrically than base loading, but makes the mechanical design of the antenna more difficult as it has to support the weight of the loading coil.

Similarly, antennas that are slightly longer than required for resonance will have some inductive reactance, which can be tuned out using a series capacitor, but this is less common.

Loop Antennas

A full-wavelength loop is another simple antennas. Loops come in all shapes: squares, rectangles and triangles to name a few. The diagram below shows a "quad loop", which has a square shape. All sides are ¼ wavelength.

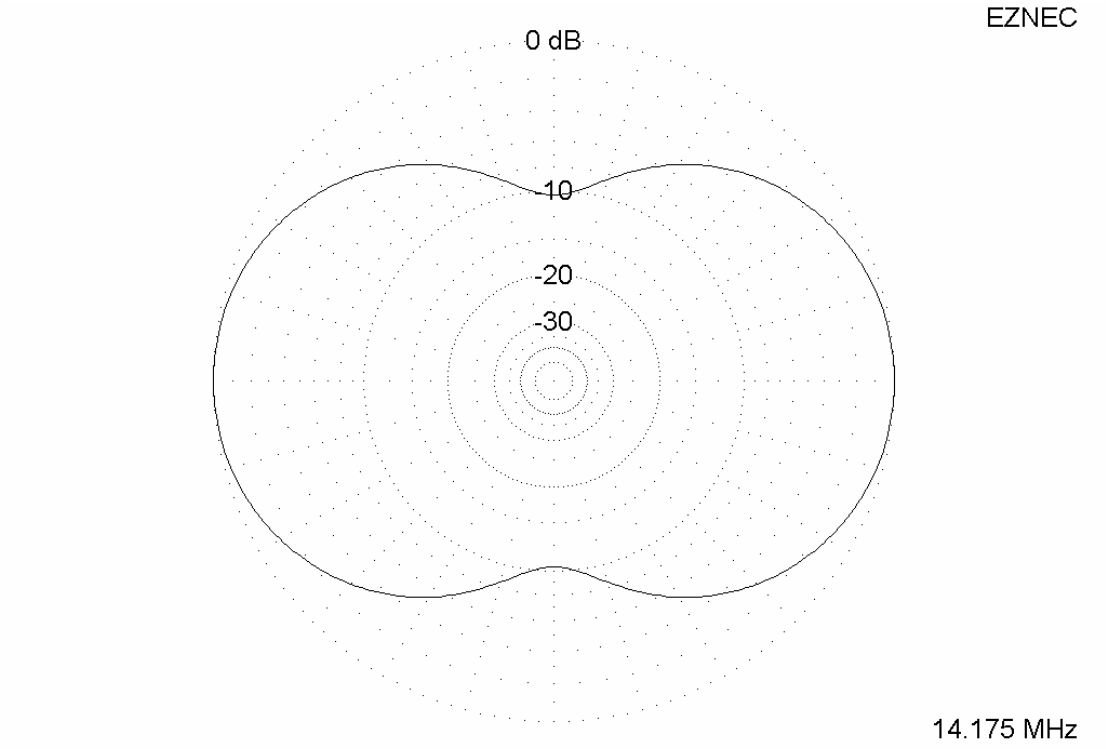


A Quad Loop

Each side of the loop is ¼ wavelength, so the total length of the loop is 1 wavelength. In loops there is no end of the antenna for the wave to reflect from, so it just keeps on going around the loop until it arrives back at the feed-point still traveling in the same direction. If the length of the loop is 1 wavelength, then the applied voltage will once again be in the same direction, so the wave traveling around the loop is reinforced, giving it a relatively low and purely resistive feed-point impedance. The radiation resistance of a loop is a bit more than a dipole, since there is more wire to radiate electromagnetic waves (remember that the radiation resistance is the apparent resistance caused by the radiation of electromagnetic waves from the antenna), typically about 130 Ω.

The loop is normally erected vertically. There are two points of high current in the loop – one at the feed-point, and one half way around the loop. If the feed-point of the loop is in the middle of one of the horizontal wires, as shown in the diagram, then both points of high current will be carrying horizontal currents and the resulting radiation will be horizontally polarized. If on the other hand the loop was fed in the middle of one of the vertical wires, then the points of maximum current would be carrying currents flowing in a vertical direction and the resulting radiation would be vertically polarized.

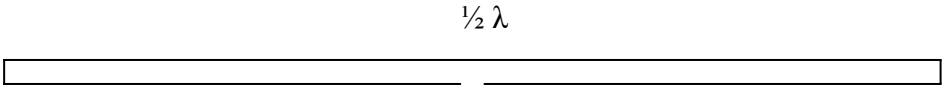
The radiation pattern of a horizontally polarized one-wavelength loop is similar to that of a dipole, with the strongest radiation being perpendicular to the horizontal wires of the loop. This is to be expected since the loop is effectively two horizontal dipoles vertically spaced by ¼ wavelength.



Azimuth pattern of a horizontally polarized quad loop

Folded Dipole

The folded dipole is a one-wavelength loop that has been "flattened" into a shape similar to a dipole.



A Folded Dipole

Its radiation pattern is the same as that of a dipole, but the radiation resistance of a folded dipole is about 300 Ω. It is higher than a normal dipole because there is more wire radiating.

Multi-element arrays

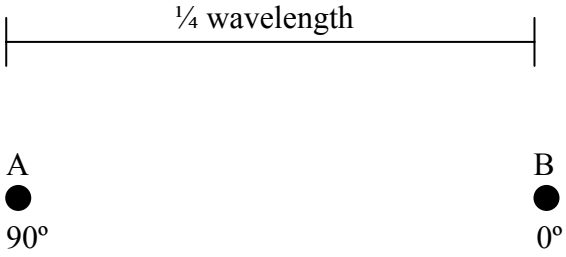
So far the antennas we have considered have all had a single radiating element. These antennas are practical and very simple to build, but have limited *directivity* – that is, the ability to favour one direction at the expense of the others. Single-element vertical antennas are *omni-directional*, radiating equally in all directions, while the dipole and quad loop are *bi-directional*, favouring two directions at the expense of the others.

However if we know the location of the radio station we want to contact, and are able to point our antenna in that direction, then it would be better to have an antenna that could direct as much of our radio waves as possible towards this station, without radiating it in unwanted directions. This would be a *unidirectional* antenna, one that favours a single direction.

It turns out that in order to make a unidirectional antenna we need at least two radiating elements, which radiate out of phase. This can be achieved in two ways. In *driven arrays*, the elements are all driven (that is, power is applied to them) in the correct phase relationship by

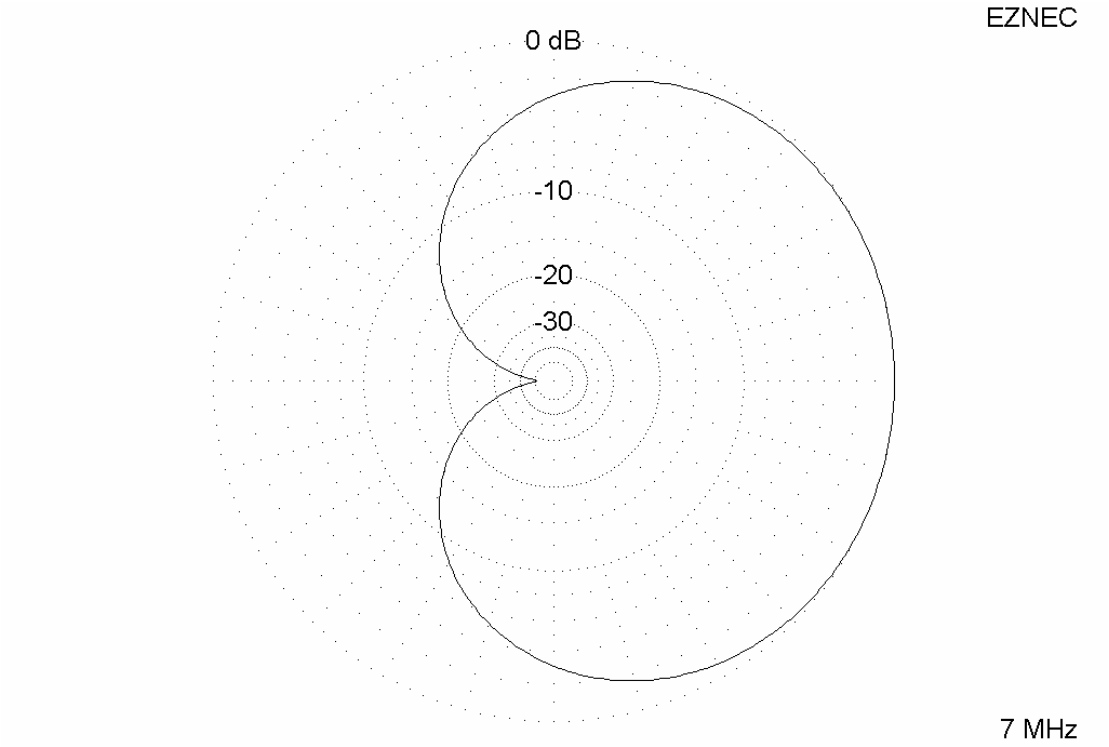
a phasing network. In *parasitic arrays*, only one element is driven, but the antenna is designed so the driven element induces currents into the other *parasitic* elements.

A simple driven array consists of two vertical elements spaced ¼ wavelength apart and driven 90º out of phase. The following diagram shows a bird's eye view of this array as seen from above.



In this case element A (the left-hand element) is shown *leading* element B (the right-hand element) by 90º; conversely element B *lags* element A by 90º. Consider radio waves leaving B and heading in the direction of A. Since the elements are spaced ¼ wavelength apart, it will take the radio waves ¼ of a cycle to get from B to A, by which time the phase of element A will have advanced by another ¼ cycle, i.e. another 90º, in addition to the 90º phasing difference that already exists between the elements. So by the time radio waves from B get to A, they are 180º out of phase with the radio waves leaving A in the same direction (i.e. from right to left) and will cancel them out.

However consider radio waves leaving A headed towards B. By the time they get to B, the phase of B will have advanced by 90º, making up for the 90º phase lag of the signal driving element B. So when radio waves from A reach B, they will be *in phase* with the radio waves radiating from element *B*, and so they will reinforce each other in the direction from left to right in the diagram. So waves heading from right to left will be cancelled; waves heading from left to right will be reinforced; and waves in different directions will be partially cancelled or partially reinforced according to some trigonometry that is too complex to go into here. The resulting pattern is shown below.



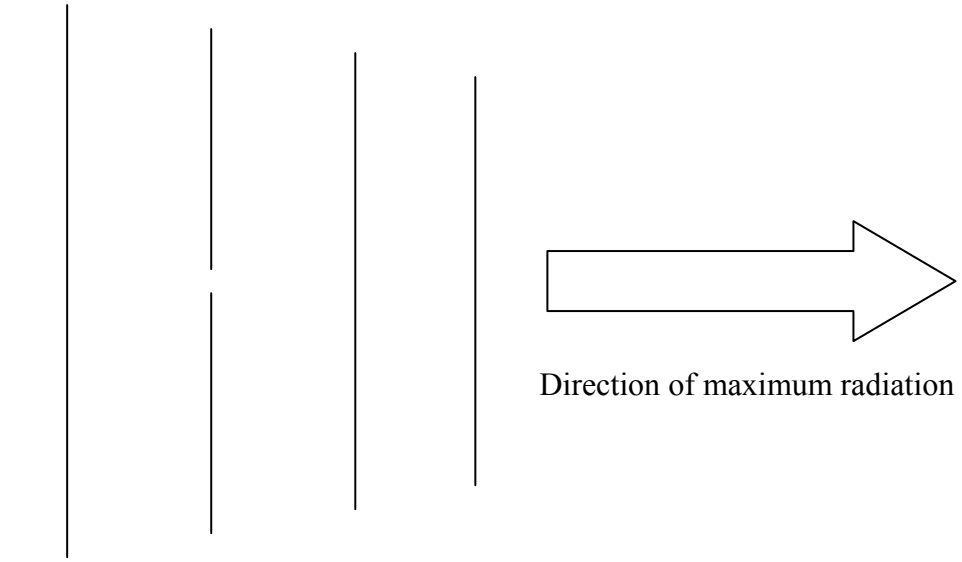
This particular shape is called a "cardioid" and it is an example of a unidirectional radiation pattern. Note the excellent null of radiation from right to left, where the signals from the two antennas exactly cancel each other.

Although the driven array seems like a simple way of getting a unidirectional pattern, there are some complications in practice. In particular, it is not as simple as it looks to generate the necessary 90º phase difference between the two signals, since the antenna elements they are driving will have different impedances due to the interaction between them. However they do have the advantage that by simply changing the phase relationship between the elements, the direction of the pattern can be reversed without having to physically rotate the antenna.

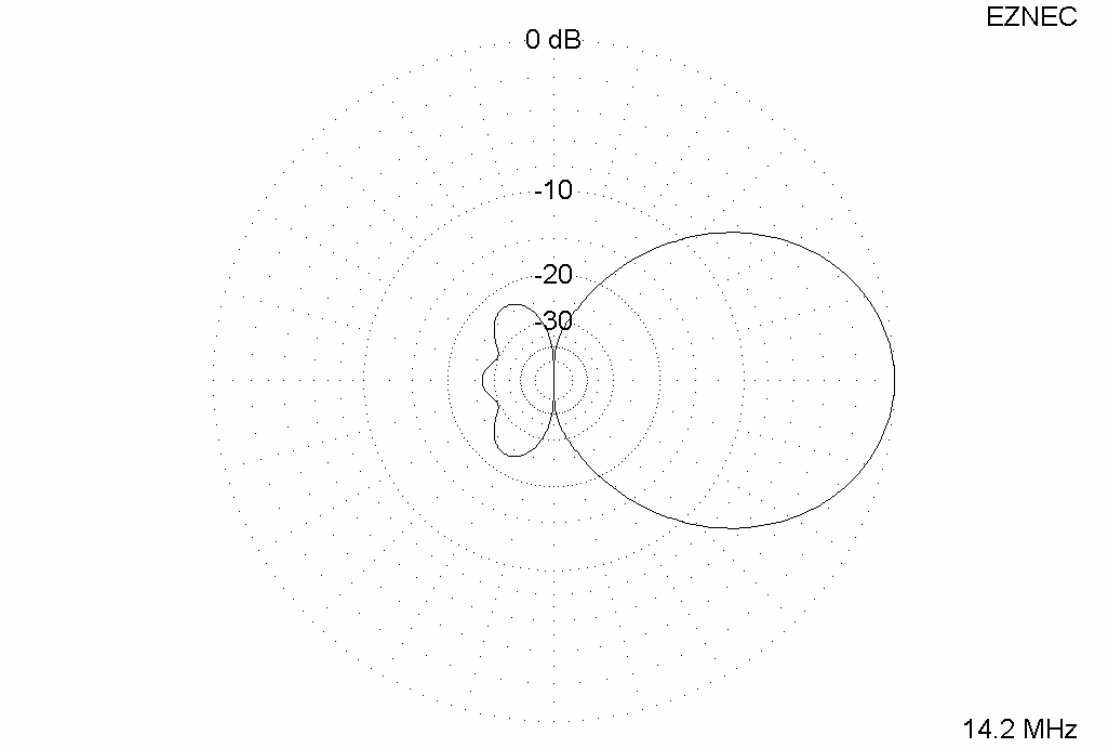
The Yagi

The full name for this antenna is the "Yagi-Uda Array", and it is named after Yagi and his supervisor Uda who invented it in 1926. But history has been a bit unkind to Professor Uda, and the antenna is normally referred to just as the "Yagi".

The Yagi consists of two or more half wavelength dipole elements, one of which – the *driven element* – is connected to the transmitter. The other element or elements are *parasitic*, meaning that currents are induced in them by induction from other elements. One of the parasitic elements will usually be a *reflector*, meaning an element that is on the other side of the driven element from the desired direction of radiation, and the other elements (if any) will be directors, meaning elements placed in the desired direction of radiation. The reflector is usually slightly longer than the driven element in order to get the correct phasing, while directors are usually somewhat shorter than the driven element. The layout of a typical 4-element Yagi is shown below:



The gap in the middle of the driven element is the feed-point, where the transmission line from the transmitter would be connected. The phasing between the elements is arranged by a careful choice of element lengths and separations so that radiation is reinforced in the desired direction and cancelled in other directions. The radiation pattern of a five-element Yagi is shown below.



Note how directional the pattern is – that is, how much radiation goes in the desired direction (from left to right) as opposed to other directions.

This makes it an excellent antenna for amateur use, since as much radiation as possible can be beamed towards a distant receiver. Yagis are popular with amateurs on the higher HF bands - especially the 20, 15 and 10 metre bands - and above. They are usually mounted on towers

with an electrically operated rotator that can point them in any desired direction. On lower bands, the large size can be a problem.

Other element configurations can also be used to make driven or parasitic arrays. For example, the cubical quad – usually called just the "quad" – is a parasitic array consisting of two or more quad loop elements. Directional antennas – including Yagis and quads – are often called "beam" antennas.

Antenna Gain

Antennas do not amplify (that is, increase the power of) signals. However a directional antenna like a Yagi will radiate more of the available power in a particular desired direction. We speak about the *gain* of an antenna to mean the extent to which it is able to concentrate its radiation in a particular direction. Gain is expressed in decibels, and is a measure of how much more power the antenna puts out in its most favoured direction, compared to the amount of energy that a reference antenna would put out in its most favoured direction.

Two different references are commonly used. The one is the half-wave dipole, in which case the unit of gain is dBd (decibels with reference to a dipole). For example, if a Yagi was found to radiate four times as much power in its favoured direction as a dipole did (given equal input powers of course), then it would have a gain of 6 dBd.

The other reference is the *isotropic radiator*. This is an imaginary antenna that radiates equally in all directions. Not just in all horizontal directions (that would be an *omni-directional* antenna like a vertical) but all directions, including up and down. Although it would be very difficult to construct such an antenna, it is easy to calculate what the strength of its radiation would be if it was constructed, so by measuring or the strength of the radiation from an actual antenna this can be compared with the calculated radiation strength from an isotropic antenna, giving a gain figure in the units dBi (dB with respect to an isotropic antenna).

In free space (that is, ignoring ground reflections) a dipole has a gain of 2,1 dBi. So if you have a gain figure for an antenna in free space in dBd, you can convert it to dBi by adding 2,1 dB. However if ground reflections are taken into account, then the gain of a dipole may be as much as 8,2 dBi. So be very careful in interpreting gain figures expressed in dBd, you need to know whether the reference dipole is at the same height as the antenna over the same ground medium, or if it is in free space.

Antenna directivity gives gain when receiving as well as when transmitting. An antenna with a gain of 3 dBd will convert signals from the desired direction into twice as much electrical power at the antenna terminals as a dipole would. This effect is not particularly useful by itself, since the limiting factor in receiver performance is usually atmospheric noise. However directional antennas also do not respond to noise coming from directions other than the favoured directions, and this reduction in noise (both manmade interference and naturally occurring noise) gives directional antennas a significant advantage when receiving.

Be very wary of accepting published figures for antenna gain. It is very difficult to measure the actual gain of an antenna, and manufacturers know that purchase decisions are often based on the gain of an antenna, so they often use devious tricks to inflate the gain figures of their antennas. Commercial manufacturers are not solely to blame and designs published in amateur radio publications may also suffer from over-optimistic gain figures. Also be suspicious of anyone who quotes an antenna gain figure in "dB" without specifying the reference antenna. A number in decibels without a reference is meaningless, and anyone who does not understand this should not be considered an authoritative source for information about antenna gain!

Effective Isotropic Radiated Power

Assume that a dipole and a Yagi are both oriented so maximum radiation is in the direction of the same receiver, and that the Yagi has a gain of 6 dBd. Then if 100 W is applied to the dipole, and only 25 W to the Yagi, the signal strength in the desired direction will be identical because the gain of the Yagi will compensate for its lower input power.

Sometimes it is useful to express the amount of power radiated in the desired direction, irrespective of the actual power input or the gain of the antenna. This can be expressed as the Effective Isotropic Radiated Power (EIRP), which is the power that you would need to supply to an isotropic antenna in order to radiate that much energy in the desired direction. EIRP can be calculated by multiplying the power actually applied to the antenna by the antenna gain with reference to an isotropic radiator.

For example, suppose our Yagi has a gain of 13 dBi, i.e. a gain of 20 with respect to an isotropic radiator, and the input power is 25 W. Then the EIRP = 20 * 25 W = 500 W. This means that to get the same amount of radiation in the favoured direction from an isotropic antenna, you would have to supply it with 500 W.

EIRP can be used to specify the power that you need to work a certain path. For example, the power required to operate a satellite might be specified as 1 000 W EIRP. You could obtain this by putting 400 W into an antenna with a gain of 4 dBi, or 100 W into antenna with a gain of 10 dBi, or 10 W into an antenna with a gain of 20 dBi.

Efficiency

The efficiency of an antenna is the amount of power radiated (in any direction) as a percentage of the amount of power supplied to the antenna. So if 100 W is supplied to an antenna, but only 40 W is radiated as radio waves, then the efficiency is 40%. The remaining energy is dissipated as heat in the antenna elements or the earth around the antenna.

Since the radiation resistance of an element drops rapidly as the length of the element is reduced below ¼ wavelength (for a vertical) or ½ wavelength (for a dipole), antennas that are considerably shorter than these standard lengths may also be very inefficient. This does not mean that it won't work – radio propagation is such that often very little radiated power is required to make a contact – but an inefficient antenna will not perform adequately when conditions are poor.

Multi-band Antennas

It is often useful to be able to use one antenna on several bands. With 8 band allocations in the HF region alone (3 MHz to 30 MHz), very few amateurs are able to put up a separate antenna for every band. There are two major considerations with multi-band antennas: radiation pattern and impedance matching.

This section will not consider the radiation pattern of multi-band antennas in any detail, except to note that the patterns may be very different on the different bands and that this is an important consideration when designing multi-band directional antennas.

Amateur transceivers are generally designed for an impedance of 50 Ω, and coax cables usually also have an impedance of 50 Ω, so the ideal antenna would also have an impedance of close to 50 Ω, or one that could be easily matched to 50 Ω. This is not hard to achieve with single-band antennas. The impedance of a dipole is around 72 Ω, and a vertical about 36 Ω, both of which are close enough to be directly fed by 50 Ω coax with no problems. However on other bands the impedance may be considerably different. For example, if a 20 m dipole is

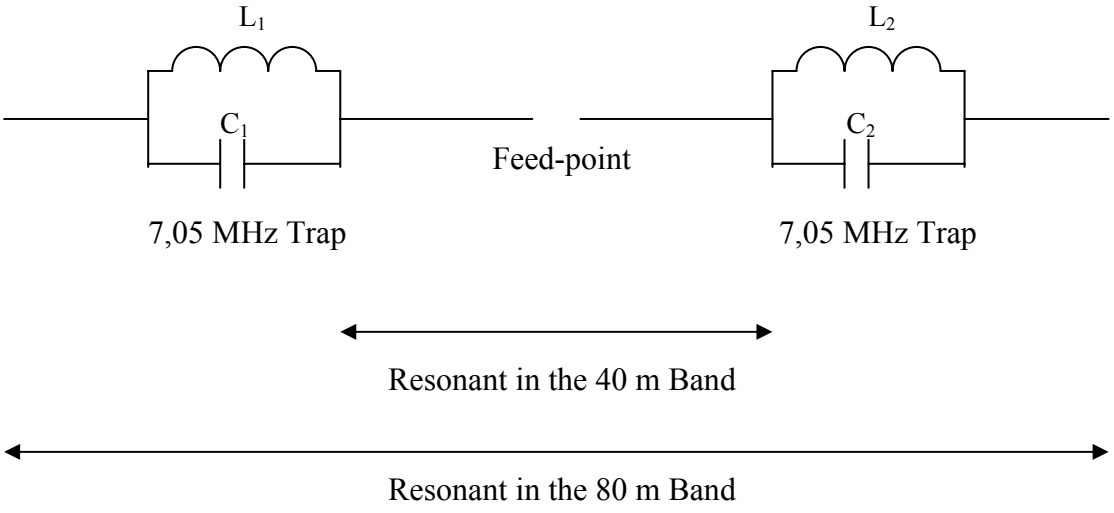
used on 10 m, the impedance will be more like 5 000 Ω, which would not be a good match for 50 Ω coax!

There are four common solutions to the impedance matching problem.

First, you can use an Antenna Tuning Unit (ATU) – also known as an "antenna coupler" or "Transmatch" - to match just about any impedance to 50 Ω. So you can use a simple single-element antenna like a dipole, and just match it on any frequency of interest using the ATU. For this to work efficiently, either the ATU must be mounted close to the antenna, or a low-loss high-impedance feeder (open wire line) must be used to connect the antenna to the ATU. You cannot expect any degree of efficiency if you run 50 m of 50 Ω coax to the antenna with the ATU on the transceiver side (i.e. not on the antenna side of the coax connection). Note that the term ATU, while frequently used, is a misnomer as the function of this unit is to match and not to tune.

Second, you can wire several elements in parallel provided that for each frequency band, one and only one element has a low, purely resistive, impedance. For example, three dipoles cut for the 20, 15 and 10-metre bands (14, 21 and 28 MHz) could be connected in parallel (i.e. all share the same feed-point). If this antenna system is fed with a 14 MHz signal, then only the 20 m dipole will have a low impedance, so almost all the energy will go into the 20 m dipole. Similarly if it is fed at 21 or 28 MHz, then only the dipole for that frequency band will have low impedance, and almost all the energy will go to that one.

Third, you can use traps to effectively shorten the antenna at higher frequencies. A trap is simply a parallel tuned circuit that appears as a high impedance at its resonant frequency, and as a low impedance at other frequencies. The diagram below shows a trap antenna for use on the 80 m and 40 m bands.



A Trap Dipole for 40 m and 80 m

The inner section of the antenna would be resonant in the 40 m band (7,0-7,1 MHz). The traps are also resonant in this frequency range, so they present a high impedance and effectively disconnect the outer sections of the antenna at these frequencies. However when the antenna is fed at a frequency in the 80 m band, the traps are low impedance so they allow the whole antenna to be active, and the total length is designed to be resonant in the 80 m band. On the 80 m band the traps have some inductive reactance, which acts like a loading coil. This means

that the total length of the antenna is less than would be required for a trapless (single band) 80 m antenna.

Trap antennas can be designed for more than two bands by adding another trap (or pair of traps in the case of a dipole) for each additional band. However since each trap will have some loss, a trap antenna for many bands might be quite inefficient on the lower bands, where antenna current is flowing through multiple traps.

Finally, you can make use of the fact that most antennas are naturally resonant on more than one frequency. For example, open-ended antennas like dipoles and verticals are resonant on odd harmonics of the fundamental frequency – that is, 3, 5 and 7 times the original design frequency. For example, a dipole designed for the 40 m (7 MHz) band may also be resonant on the 15 m (21 MHz) band, which is the third harmonic of the design frequency. Full-wavelength closed loop antennas like the quad antenna are resonant on all harmonics, so in theory an 80 m quad loop should also be resonant in the 40, 20 and 10-metre bands. While this sounds great, in practice you may find that the resonant frequencies are not precisely where you need them, and that it is difficult to tune them to the desired frequency without changing the other resonant frequencies!

Summary

Antennas convert electrical energy into radio waves that can be radiated long distances. Electromagnetic waves consist of a magnetic field and an electric field that are both at right angles to each other, and at right angles to the direction of propagation of the wave. The *polarization* of radio waves depends on the orientation of the electric field – if the electric field is horizontal, the wave is said to be *horizontally polarized* and if it is vertical, the wave is *vertically polarized*. Antennas respond less to radio signals with the "wrong" polarization.

The half-wave dipole consists of a centre-fed ½ wavelength element. Its radiation resistance is approximately 72 Ω and a horizontal dipole over ground has a bi-directional pattern similar to a figure "8". The ¼ wave vertical consists of a vertical ¼ wavelength element that is fed either against ground or against a "ground-plane". It has an omni-directional pattern, radiating equally in all (horizontal) directions.

A unidirectional pattern can be achieved using a multi-element antenna, either a *phased array* where each element is individually fed from a phasing network, or a *parasitic array*, where the transmitter feeds only one element and the others are excited by inductive coupling from other elements. The most common parasitic array is the Yagi, which consists of two or more dipole elements – the driven element, a reflector, and one or more directors.

The gain of antenna expresses how much power is radiated in the most favoured direction, compared with some reference antenna. Gain can be specified in dBd (dB with reference to a dipole) or in dBi (dB with reference to an isotropic radiator). The efficiency of antenna is the amount of power radiated as a percentage of the total power applied to the antenna. The Effective Isotropic Radiated Power (EIRP) is the power fed to the antenna multiplied by the gain of the antenna with respect to an isotropic radiator.

An antenna may be impedance-matched on multiple bands by using an antenna tuner, by feeding multiple elements in parallel, by using traps or by taking advantage of naturally occurring harmonic resonances.

Revision Questions

1 Electromagnetic waves are created by:
a. The alternating RF currents in an antenna.
b. Magnetic solenoids.
c. Audio loudspeakers.
d. DC voltages.

2 In electromagnetic radiation, which of the following is true?
a. E and H are at 180° to each other.
b. E, H and the direction of propagation are all at right angles to each other.
c. The angle between E and H is 0 °.
d. The velocity of propagation is at 180° to the E field but in line with the H field.

3 In order to radiate, an electromagnetic wave must have:
a. E Field only.
b. H Field only.
c. E and H Field.
d. Air to travel in.

4 Polarization of an electromagnetic wave is fixed by:
a. The direction of the H field.
b. The direction of propagation.
c. By an anti-phase signal.
d. The orientation of the transmitting antenna.

5 The wavelength of a signal of 100 MHz in free space is:
a. 30 mm.
b. 0,3 m.
c. 3,0 m.
d. 30,00 m.

6 When the resonant length of an antenna matches the transmitted frequency:
a. Maximum power will be reflected.
b. A good SWR will be obtained.
c. The SWR will be poor.
d. An SWR reading will be meaningless.

7 What do the terms vertical and horizontal, as applied to wave polarization, refer to?
a. Orientation of the electric lines of force.
b. Orientation of the magnetic lines of force.
c. Orientation of the charge particles in the propagation medium.
d. Launching angle of the wave with respect to the earth's surface.

8 What radiation pattern does an ideal half-wave dipole have if it is installed parallel to the earth?
a. It radiates well in both directions parallel to the earth and perpendicular to the dipole.
b. It radiates poorly in directions parallel to the earth and parallel to the dipole.
c. It radiates equally well in all directions parallel to the earth.
d. It radiates poorly in all directions parallel to the earth, but it radiates well in directions perpendicular to the earth.

9 How does proximity to the ground affect the radiation pattern of a horizontal dipole antenna?

a. If the antenna is too far from the ground, the pattern becomes unpredictable.
b. If the antenna is less than one-half wavelength from the ground, reflected radio waves from the ground distort the radiation pattern of the antenna.
c. A dipole antenna's radiation pattern is unaffected by its distance to the ground.
d. If the antenna is less than one-half wavelength from the ground, radiation off the ends of the wire is reduced.

10 Which kind of antenna would be least affected by signal emanating from a particular direction, enhancing the signals from a desired direction?

a. A monopole antenna.
b. An isotropic antenna.
c. A vertical antenna.
d. A beam antenna.

11 What is a directional antenna?

a. An antenna whose parasitic elements are all constructed to be directors.
b. An antenna that radiates in direct line-of-sight propagation, but not skywave or skip propagation.
c. An antenna permanently mounted so as to radiate in only one direction.
d. An antenna that radiates more strongly in some directions than others.

12 What is the purpose of an antenna matching circuit?

a. To measure the impedance of the antenna.
b. To compare the radiation patterns of two antennas.
c. To measure the SWR of an antenna.
d. To match impedances within the antenna systems.

13 When will a power source deliver maximum output?

a. When the impedance of the load is equal to the impedance of the source.
b. When the SWR has reached a maximum value.
c. When the power supply fuse rating equals the primary winding current.
d. When air wound transformers are used instead of iron core transformers.

14 What is a Yagi antenna?

a. Half-wavelength elements stacked vertically and exited in phase.
b. Quarter-wavelength elements arranged horizontally and excited out of phase.
c. Half-wavelength linear driven element(s) with parasitically excited parallel linear elements.
d. Quarter-wavelength, triangular loop elements.

15 Why is a Yagi antenna often used for amateur radio communications on the 20 meter band?

a. It provides excellent omni directional coverage in the horizontal plane.
b. It is smaller, less expensive and easier to erect than a dipole or vertical antenna.
c. It discriminates against interference from other stations off to the side or behind.
d. It provides the highest possible angle of radiation for the HF bands.

16 Choose a physical description of the radiating elements of a horizontally-polarized Yagi antenna.

a. Two or more straight, parallel elements arranged in the same horizontal plane.
b. Vertically stacked square or circular loops arranged in parallel horizontal planes.
c. Two or more wire loops arranged in parallel vertical planes.
d. A vertical radiator arranged in the centre of an effective RF ground plane.

17 What is the name of the parasitic beam antenna using two or more straight metal-tubing elements arranged physically parallel to each other?
a. A quad antenna.
b. A delta loop antenna.
c. A zepp antenna.
d. A Yagi antenna.

18 How many driven elements does a Yagi antenna have?
a. None; they are all parasitic.
b. One.
c. Two.
d. All elements are driven.

19 What kind of antenna array is composed of a square or diamond-shaped full-wave closed loop driven element with parallel parasitic element(s)?
a. Dual rhombic.
b. Cubical quad.
c. Stacked yagi.
d. Delta loop.

20 What is the polarization of the signal from a half-wavelength antenna which has elements perpendicular to the earth's surface?
a. Circularly polarized waves.
b. Horizontally polarized waves.
c. Parabolically polarized waves.
d. Vertically polarized waves.

21 A two trap dipole will allow operation on:
a. One band.
b. All bands.
c. Three bands.
d. Two bands.

22 A folded dipole has an approximate impedance of:
a. 50 Ω.
b. 72 Ω.
c. 150 Ω.
d. 300 Ω.

23 A vertical antenna relies upon :
a. A good earth and ground connection.
b. No earthing.
c. A sensitive receiver.
d. The D layer.

24 The term Zepp, Yagi, Quad and Log Periodic refer to:
a. Oscillators.
b. Transistors.
c. Antennas.
d. Diodes.

Chapter 26 - Propagation

Propagation means the process by which radio waves get from the antenna of the transmitter to the antenna of a distant receiver. This chapter introduces the different propagation modes used by amateurs.

Direct Wave (line of Sight) Propagation

Electromagnetic radiation generally travels in straight lines, so if radio waves can travel straight from the transmitting antenna to the receiving antenna without passing through anything that blocks it, then communication is possible. This simplest form of propagation is known as "direct wave" propagation. It is also called "line of sight" propagation although this term is a bit misleading since some things that block light, such as a wooden structure, are transparent to radio waves.

Direct wave propagation affects all frequencies. The range possible depends on the terrain and the height of the antennas. In flat terrain, with both antennas 10 m high, the range of direct wave propagation is about 20 km. However hilly terrain can be used to good effect by placing one of the antennas on top of a hill where it can be "seen" from much further away. This is why VHF repeaters are usually located on high sites, since they rely on direct wave propagation.

Both horizontally polarized and vertically polarized waves propagate equally well over line of sight. However because this form of propagation retains the original polarization of the wave, it is important to ensure that both transmitting and receiving antennas have the same polarization.

Ground Wave Propagation

Low and medium frequencies *refract* around the surface of the earth. Refraction is caused by the nearby ground slowing the radio wave down slightly, causing it to bend towards the ground. However because the ground itself is bending with the curvature of the earth, the effect is that the ground wave "hugs" the earth. Refraction is most pronounced at lower frequencies, so this effect is most significant in the low and medium frequency bands. It is present but less effective in the high frequency (HF) bands and absent at VHF and above.

This is why you can still listen to medium wave AM commercial broadcast stations up to 100 km or so away from the transmitter – the medium wave broadcast frequencies (530 kHz – 1.6 MHz) are low enough for good ground wave propagation to occur. However commercial FM transmitters use VHF frequencies (88-108 MHz), which are only propagated by direct wave, so they are only usable within 10 or 20 km of the transmitter.

The same ground interactions that allow the wave to refract around the curvature of the earth also attenuate it, limiting the range of ground wave propagation to a few hundred kilometres, depending on the power of the transmitter.

The Atmosphere

The atmosphere consists of three layers: troposphere, stratosphere and ionosphere. The troposphere extends from the surface of the earth to a height of about 10 km. It is the area where most of the weather we are familiar with happens. The stratosphere extends from 10 km above the surface of the earth to approximately 50 km. In this region the temperature and humidity remain relatively constant, and it has little effect on propagation.

The ionosphere is that part of the upper atmosphere where free electrons occur in sufficient density to have an appreciable influence on the propagation of radio frequency

electromagnetic waves. It extends from approximately 50 km to 800 km above the earth's surface. In the ionosphere, high-energy solar radiation (x-rays, ultraviolet radiation and particles from the "solar wind") strips electrons from some gas molecules, leaving positively charged ions and free electrons.

The ionosphere is divided into four layers. The D layer, which extends from 50-90 km above the surface, is only present during daylight hours. As soon as the sun's ionising radiation is no longer present, electrons and ions rapidly recombine to form neutral (un-ionised) gas and the D layer disappears. The principle effect of the D layer is to absorb radio waves. Although some absorption takes place at all frequencies, the amount of absorption decreases with the square of the frequency, so it affects low frequencies much more than high frequencies.

The upper three layers have a different effect. Instead of simply absorbing radio waves, they bend them by refraction. If a wave is bent sufficiently then it may return to earth a considerable distance from the transmitter, almost as though it had been reflected off the ionosphere. The amount of refraction (bending) depends on frequency and is more pronounced at lower frequencies. The upper layers are the E layer, which extends from 90 to 150 km above the surface; the F1 layer, which extends from 150 to 180 km; and the F2 layer, which extends from 180 km to 300 km or higher. At night the E layer dissipates while the F1 and F2 layers combine to form a single F layer that is less strongly ionised than during the daytime.

Sky Wave (Ionospheric) Propagation

The effect of this is that during the daytime, the D layer will absorb low frequencies, but higher frequencies will make it through the D layer (albeit with some attenuation) and can be refracted back to earth by the E, F1 or F2 layers. Higher frequencies still will not be refracted sufficiently by the E, F1 and F2 layers and will continue out into space instead of being returned to earth. Refraction by the F2 layer (or at night by the single F layer) is responsible for most long-distance HF communication.

At night, the D layer dissipates almost immediately and the E layer more gradually, while the F1 and F2 layers combine to form a single less strongly ionized layer. Now that there is no D layer to absorb low frequencies, they can be reflected from the F layer and travel long distances. However high frequencies are not being refracted sufficiently by the more weakly ionized nighttime F layer, so they will be lost into space.

Note that fairly high frequencies may still be usable well after local sunset. This is because it takes a considerable time for ionization levels in the F1 and F2 layers to decrease to their nighttime values. Also, because these layers are high above the surface of the earth, they will be illuminated for some time after local sunset. And finally, for paths from east to west, the point where the waves are refracted will be located some distance to the west of the transmitter, where sunset would have been later.

This process of being refracted from the ionosphere is also known as "skip". The maximum skip distance for the E layer is around 2 500 km, and about 5 000 km for the F layer. Longer paths may be achieved by multi-hop propagation, where the refracted signal bounces of the surface of the earth back to the ionosphere and is refracted back to earth again.

The highest frequency that can be used on a particular path (i.e. for communication between two particular places at a particular time) is called the Maximum Usable Frequency (MUF) for that path. It is dependent mainly on the degree of ionization present in the ionosphere. Increasing the effective radiated power (ERP) of the signal won't help, because if the frequency is above the path MUF then the additional power will just be radiated into space.

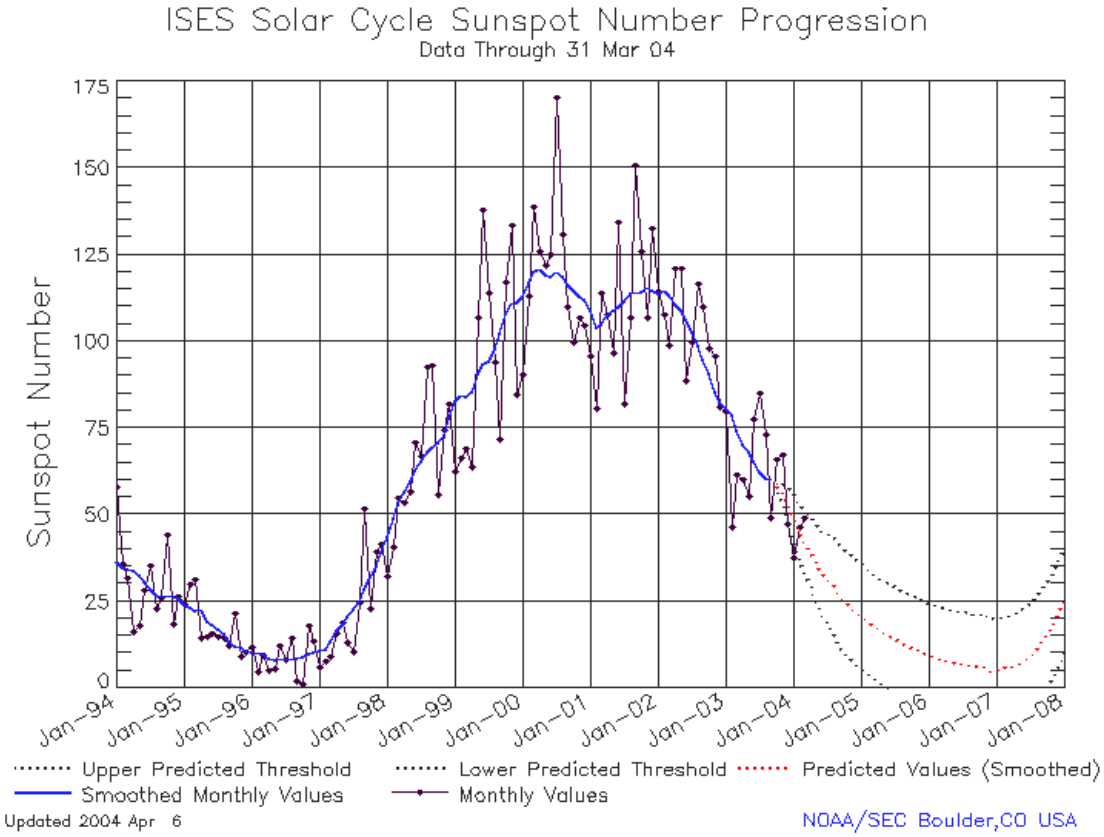
The lowest frequency that can be used for a path is called the Least Usable Frequency (LUF). The LUF depends not only on the amount of ionization, but also on the amount of atmospheric and man-made noise present at the receiver and on the ERP of the signal. This is because the main consideration is whether the transmitted signal, after being attenuated by the D layer, is still sufficiently strong to be heard above the noise, so additional power will help.

D-layer absorption and atmospheric noise both increase as the frequency decreases, so the best propagation is usually to be found just below the MUF.

The critical frequency is the highest frequency that can be radiated vertically upwards and still return to earth. Note that the MUF for a path may be well above the critical frequency, since waves radiated at a shallow angle may be returned to earth when waves radiated vertically upwards are not. The critical frequency is only of indirect interest to amateurs, since usually one is not hoping to bounce a signal vertically off the ionosphere and have it return to the house next door. However the critical frequency does give a general idea of ionospheric conditions, so when it is high then path MUFs are likely to be high as well; when it is low, path MUFs will be low.

MUFs and LUFs depend on the extent of ionization in the ionosphere. This varies with the time of day; with the season and with the amount of solar activity. Solar activity follows cycles approximately 12 years long. Every 12 years there is a solar maximum, when high levels of solar activity generate intense ionization and MUFs can extend well above 50 MHz during daylight hours, giving great openings on the 6 m and 10 m bands. About 6 years later there will be a solar minimum, where MUFs may be under 20 MHz. The 6m and 10 m bands will be virtually dead, while propagation on the low bands (160 m and 80 m) will be better than usual.

Solar activity is measured in two different units: the *sunspot number* and the *solar flux index* (SFI). The following graph shows the progression of the current solar cycle.



The sunspot cycle affects the optimum frequencies for communication. At the sunspot maximum, daylight communication will typically use the 15 m, 12 m, 10 m and 6 m amateur bands, with the other bands being used only at night. At a sunspot minimum, daytime bands will typically be 40 m, 30 m and 20 m, with 160 m and 80 m being used at night.

Of course propagation does not depend only on conditions at the transmitter, but equally on conditions at the receiver and along the whole path. For example, it would not make much sense to try to contact stations in the USA on a daylight band at 10 am local time in South Africa, since at that time it is only 3 am local time on the east coast of the USA. However at 10 am local time in South Africa daylight bands might work well for contacting Japanese stations, since then it is 5 pm in Tokyo. In general, paths that are either all daytime or all-night are the easiest. Mixed daylight and nighttime paths can be difficult since frequencies that work on one side of the link won't work on the other side, and vice-versa. However remember that "daytime" conditions may persist for several hours after local sunset, and even later for east-west paths.

In general radio waves travel along the great circle path between their source and destination. This is a curved path that represents the shortest distance between two points on the surface of the earth, without actually going *through* the earth! However there are *two* great circle paths between any points on earth – a short path, and a long path. For example, the short path to the USA from South Africa is to the northwest; while the long path is in the opposite direction, to the southeast, and travels the other direction around the world, finally ending up at the west coast of America. Radio propagation can be via either long or short paths, depending on the conditions along each path.

Ionospheric propagation achieves the longest distances when the takeoff angle of the radio signal (the vertical angle of the signal above the horizon) is small. This is for two reasons: a signal with a low radiation angle will hit the ionosphere further away from the transmitter than one with a high takeoff angle, resulting in a greater skip distance; and because it requires less refraction by the ionosphere to be returned to earth, higher frequencies will be usable that would be the case for a signal with a higher takeoff angle. So for long-range ("DX") communication, a low takeoff angle is desirable. In the case of horizontally polarized antennas, this means the antenna should be as high above ground as possible.

Ionospheric propagation is most common on the medium frequency bands (at night only) and high frequency bands. It occurs occasionally in the low VHF region, for example the 6 m amateur band, which is also called the "magic band" because "skip" although infrequent can travel great distances with very little power.

Sporadic E Propagation

A form of ionospheric propagation that affects VHF transmissions is known as Sporadic-E. This consists of the refraction of VHF frequencies from small patches of intense ionization in the E layer. The cause of these intensely ionized areas is not well understood and they appear unpredictably (hence "sporadic") most frequently during summer daylight hours and may last for several hours. Sporadic E is usable on frequencies from 28 MHz to 220 MHz and signal strengths are often very strong, with low-power transmitters being heard hundreds of kilometres away. Path distances may exceed the 2 500 km maximum for single-hop E layer skip, indicating that either multiple hops or some form of ionospheric ducting is present.

Meteor Scatter

Meteors entering the earth's atmosphere leave a trail of ionized gas in their path that can refract VHF signals. The ionization typically only lasts for a few seconds to tens of seconds before the electrons and ions recombine. This means that specialized digital modes like JT6M

and FSK441 are needed to take advantage of meteor scatter. Typical meteor scatter bands are 6 m and 2 m.

Tropospheric Bending, Scatter and Ducting

Some bending (refraction) of VHF and UHF radio waves occur in the troposphere, which increases the "radio horizon" (the distance over which radio waves can propagate without reflection or scattering) by about 15% compared to the visual horizon.

Temperature and humidity irregularities within the troposphere (the lower 10 km of the atmosphere) can reflect VHF and UHF signals over a distance of from 100 to 500 km or so. The reflections are usually fairly weak, so reasonable ERP (either a high power transmitter or an antenna with gain) is required. However unlike meteor scatter, tropo-scatter is long lived, so it is possible to use standard modes like CW and SSB for tropo-scatter work. FM is not recommended, as it requires more power for an intelligible signal than either CW or SSB.

In tropospheric ducting, VHF signals are "trapped" between an inversion layer and the ground or between two inversion layers and may travel thousands of kilometres with little attenuation.

Earth Moon Earth (EME) and Satellite

EME consists of bouncing signals off the moon to some distant location on earth. It used to be the exclusive preserve of those with very high power stations and large steerable antenna arrays; but the new weak-signal digital modes like JT65 mean that today even modestly equipped stations can experience EME contacts – especially if the station on the other side of the contact has high power and a large steerable antenna array!

There are a number of amateur satellites in orbit that will relay signals in various modes, including FM, SSB, CW and digital modes. They act like terrestrial repeaters, except that signals are usually sent up to the satellite on one band (the *uplink* band) and received from it on a different band (the *downlink* band). A pair of uplink and downlink bands is known as a *mode.* Modes are described using two letters, which represent the uplink and downlink bands, for example Mode V/U which has a 2 m uplink and a 70 cm downlink. (The "V" stands for "VHF" referring to the 2 m band, and the "U" for "UHF" referring to the 70 cm band).

The easiest satellites to work are those in low earth orbit, which because they are fairly close to the earth (100-200 km above the surface) can be worked with low power and simple antennas. Because they are low they offer a fairly small footprint (the area in which stations can communicate via the satellite) and short pass times – often only a few minutes. The high earth orbit satellites like AO-40 offer a larger footprint and much longer pass times, but require more sophisticated equipment to access.

Summary

Direct wave (line of sight) propagation is when signals of any frequency travel directly from the transmitter to the receiver. Ground-wave propagation is where low and medium frequency signals are refracted around the curvature of the earth, up to a distance of several hundred kilometres.

Ionospheric propagation results from the refraction of radio wave by the E, F1 and F2 layers of the ionosphere. During daylight hours, the D layer absorbs low-frequency signals, so only higher frequencies are usable. The D layer dissipates rapidly after dark, allowing even low frequency signals to reach the F layer. High frequency signals are not refracted sufficiently by the ionosphere to return to earth, but are lost into space. The critical frequency is the highest frequency at which radiation directed vertically upwards will return to earth. The maximum

usable frequency (MUF) for a particular path is the maximum frequency that will be refracted by the ionosphere along that path and it may be considerably higher than the critical frequency. The lowest usable frequency (LUF) is the lowest frequency that can be used for communication on a particular path, and depends on the ERP of the transmitter and receiver noise level as well as the extent of ionization. Ionospheric propagation via the F layer occurs most commonly for the high frequency (HF) bands, although there are occasional openings on the 6 m band. The amount of ionization depends on the time of day, season and the twelve-year solar cycle.

Sporadic E propagation consists of the refraction of VHF signals by intensely ionized patches of the E layer. These patches occur sporadically but may last for several hours and allow VHF communication at ranges from a hundred to several thousand kilometres. Meteor scatter uses specialized digital modes to communicate using the very brief periods of intense ionization caused by meteors entering the earth's atmosphere. Tropospheric scatter results from signals being reflected by temperature and humidity differences in the troposphere and can result in consistent VHF and UHF communications over ranges of 100 to more than 500 km with suitable equipment. Tropospheric ducting, when VHF signals are trapped between the ground and an inversion layer or between two inversion layers, is much less common but can result in signals being received with good strength thousands of kilometres away.

Earth-moon-earth (EME) is possible with modest stations using weak signal digital modes. Amateur satellites retransmit signals received on one frequency band onto another frequency band, functioning similarly to repeaters on earth but over much greater distances.

Revision Questions

1 What is the propagation path of a wave that travels directly from the transmitting antenna to the receiving antenna called?

a. The ground wave.
b. The sky wave.
c. The linear wave.
d. The plane wave.

2 What effect does tropospheric bending have on 2 meter radio waves?

a. It increases the distance over which they can be propagated.
b. It decreases the distance over which they can be propagated.
c. It tends to garble 2-meter phone transmissions.
d. It reverses the sideband of 2-meter phone transmissions.

3 Two stations 5 km apart are most likely to be communicating via:

a. Tropospheric waves.
b. Ionospheric waves.
c. Ground waves.
d. Telephone.

4 The D layer occurs in the Ionosphere at:

a. 80 km.
b. 150 km.
c. 200 km.
d. 300 km.

5 The F2 layer occurs at:
a. 80 km above the earth.
b. 150 km above the earth.
c. 100 to 200 km above the earth.
d. 200 to 300 km above the earth.

6 The ionospheric layer that mostly affects long distance radio communications is:
a. D layer.
b. E layer.
c. F1 layer.
d. F2 layer.

7 Signals above the maximum usable frequency passing through the F2 layer:
a. Are reflected to earth.
b. Pass through and are lost in space.
c. Are amplified.
d. Are attenuated and refracted.

8 A VHF station finds a propagation opening on 2 m that lasts for an hour, with contacts of around 1 000 km. This is most likely caused by:
a. Sporadic E.
b. Tropospheric scatter.
c. Ionospheric refraction in the F layer.
d. Meteor scatter.

9 Meteor scatter QSOs:
a. Often use SSB.
b. Are necessarily very short.
c. Are only possible in summer.
d. Are common on the lower HF bands.

Chapter 27 - Electromagnetic Compatibility

Electromagnetic compatibility (EMC) is the process of ensuring that equipment that radiates electromagnetic radiation, such as an amateur transmitter, does not interfere with equipment that may be sensitive to electromagnetic radiation, such as television and radio receivers. There are two considerations when dealing with interference problems. The first is a legal and social one: who is responsible for solving the interference problem. (I say legal *and* social because sometimes a purely legal approach will generate quite undesirable social results). The second consideration is technical: what are the causes of interference, and how can it be eliminated.

EMC problems can be classified according to whether the device that is radiating the signal causing interference is an *intentional radiator* (that is, a device that is intended by virtue of its function to radiate, such as an amateur transmitter or a garage opening remote control) or an *unintentional radiator* (that is, a device that does not need to radiate in order to perform its intended function, such as a motor vehicle ignition system or an electric fence) and whether the device being interfered with is a *receiver* (that is, equipment designed to receive radio signals at some frequency) or is not a receiver (such as a CD player that is picking up breakthrough from nearby transmissions).

Unintentional Radiators

There are strict limits to the maximum permitted radiation from any system that does not have to radiate in order to operate correctly. If a system that does not include a radio transmitter of some kind is causing interference, then that is generally because the system is radiating more than permitted, and it should be repaired or replaced at the owner's expense.

For example, if you receive interference from a neighbour's electric fence, then that probably indicates that the electric fence is radiating more than is permitted, and the neighbour is responsible for having the defect rectified, and must turn the electric fence off until it complies with requirements. Of course convincing your neighbour of this may be difficult!

Interference to non-receiving equipment

The converse applies when the equipment being interfered with is not intended to receive radio signals. For example, suppose your neighbour reports that your radio transmissions are "breaking through" on their stereo system when they are listening to CDs. Because the stereo system when listening to CDs is not supposed to receive radio signals, the problem lies with the stereo, not with the radio transmitter.

Often the root cause is that the affected equipment was not designed for, and has not been tested in, environments with strong RF signals present. Unfortunately it is quite legal for such equipment to be sold, and it will work fine for 99% of the time, since in most locations it will encounter only weak electromagnetic radiation from distant transmitters. Then an amateur moves in next door, sets up equipment that is operating within the limits of their license, and all of a sudden the neighbour's CD player receives interference. It is quite natural for the neighbour to think that this is the amateur's fault, and that they must fix the problem or stop transmitting. However in actual fact, the fault lies with the manufacturer of the equipment for not designing it to withstand the levels of electromagnetic radiation that may result form a nearby amateur installation.

In this case, even though it is the neighbour's responsibility to solve the problem, it would be diplomatic for the amateur concerned to make his or her technical skills available to the neighbour to help diagnose the problem and suggest solutions. Apart from good neighbourliness, the same neighbour may have the opportunity to comment on your application to erect a tower, and is more likely to be kindly disposed to such a request if you

have helped them to solve any problems that appear to have been caused by your transmissions in the past!

Intentional Radiators interfering with Receivers

The situation is slightly more complex if an intentional radiator (such as your amateur transmitter) interferes with a device that is intended to receive radio signals (such s your neighbour's television). In this case, the key question is the nature of the interfering signal.

If the interfering signal is in all respects a legal licensed transmission – that is, it is within an amateur band, does not exceed the power permitted for the band and license holder, and is a clean signal – then the problem is being caused by the receiving equipment being affected by an out of band signal, and it is the receiving equipment that is defective and must be repaired.

On the other hand if the transmitted signal in any way does not conform with the requirements of your license, then you should first correct the problem with the transmitted signal before suggesting to your neighbour that they have their TV fixed! This is particularly important because if interference is reported to ICASA then their first course of action will probably be to inspect your transmitting equipment. If it is found to be out of order in any way then you may be held responsible for the interference and, even if you are not, the transmitting equipment can be confiscated if it does not conform with your license requirements.

Once again, as a matter of diplomacy, it is a good idea to assist your neighbour if possible to solve the interference problem, even if you have determined that your transmitter is operating quite legally. As well as maintaining peace in the neighbourhood, this will help to maintain the good reputation of amateur radio. However if this is not possible – for example, if your neighbour refuses your assistance and insists that you just stop operating – then as long as you are certain that your equipment is operating legally, then you are entitled to continue to operate despite the interference to your neighbour's television or other equipment.

Shared Bands

One exception to this is that some amateur bands are shared between different users, with one of the users being declared the "primary" user and the other as "secondary" users. For example, amateur radio has been allocated the 2 GHz band (2,3-2,45 GHz) on a secondary basis; the primary use is industrial, scientific and medical.

Simply put, secondary users may not cause interference to primary users (and must stop operating if this is the only way to prevent interference), while they must accept interference from primary users. So if you live next door to a hospital and receive interference from medical equipment that is intentionally radiating in the 2 GHz band, then there is nothing you can do about it.

Of course all amateur bands are shared with other amateurs, and it is important that we take steps to avoid interfering with our fellow amateurs. This should include operating courtesy and ensuring that your transmitter is radiating a clean signal.

Causes of Interference

There are three possible causes of interference.

1. The transmitter may be radiating on a frequency that it should not be radiating on.

2. The receiver might be receiving signals that it should not be.

3. The transmitter and receiver may both be working correctly, but something else is translating the transmitted signal to the frequency of the receiver. For example, corrosion can cause metal to operate like a rectifier, re-radiating harmonics of signals transmitted from a nearby transmitter.

Since the latter is quite uncommon and usually requires specialised equipment and significant expertise to resolve, we will only look at the first two possibilities.

Transmitter Defects

The most common problems in transmitters are frequency instability, harmonic radiation, spurious oscillations, and "wide" signals.

Frequency instability is usually the result of LC (inductor/capacitor) oscillators that have not been adequately compensated for temperature variations or protected against mechanical shock. It is most likely to impact on other amateurs, unless the instability is sufficient to take the transmitter out of the amateur band and cause interference to other services. Fixing frequency instability usually requires design modifications or improved construction methods (for example, more solid construction that is less sensitive to mechanical knocks). It is quite uncommon with modern crystal-controlled radios, although it may occur if a PLL frequency synthesizer gets unlocked from the reference frequency.

Another type of frequency instability is chirp, which occurs when the oscillator frequency is affected by the loading of subsequent stages or by fluctuations in the power supply voltage when a CW transmitter is keyed. It can be prevented by using a high-impedance buffer amplifier after the oscillator; and by regulating the oscillator voltage supply.

Harmonic radiation occurs on multiples of the transmitter output frequency. For example, a transmitter operating at 144 MHz may interfere with a television receiver operating at 720 MHz (720 = 144 * 5). It can be caused by overdriving an amplifier stage (for example by having the microphone gain or CW drive level set too high) or by inadequate attenuation of harmonics by the transmitter's output low-pass filter.

If the problem is caused by overdriving the transmitter, then the solution is to reduce the drive level by adjusting the microphone gain or CW drive correctly. However if the problem persists even when the transmitter is not being overdriven , then the best solution is to add an additional low-pass filter between the transmitter and the antenna. Low-pass filters for the HF bands (up to 30 MHz) are available at reasonable cost and provide substantial attenuation at higher frequencies, typically 50 dB or better at 50 MHz.

Another solution sometimes recommended is to use an antenna tuning (matching) unit (ATU) even when it is not required to match the antenna, as the ATU may attenuate out of band signals. We think this was probably better advice in the days when most ATUs used a pi configuration and also acted as low-pass filters. Today many ATUs use a T configuration, and would act as a high-pass filter, making their value for reducing harmonic radiation questionable. In any case, since this is not what the ATU is intended for, there is no guarantee that it will be effective, so my advice would be to use a purpose-built low-pass filter instead.

Spurious oscillations may either be self-oscillation, at or near the intended frequency of operation of an amplifier or mixer, or parasitic oscillations, which usually occur at VHF or UHF frequencies. Self-oscillation is caused by unintended feedback from the output of an amplifier or mixer that includes tuned circuits to its input, causing oscillation at the resonant frequency of the tuned circuit. It can be suppressed either by reducing the coupling (for example by shortening component leads) or by introducing negative feedback to reduce the loop gain and prevent oscillation.

Parasitics are VHF or UHF oscillations that occur due to unwanted "hidden" resonances in oscillators and amplifiers – for example, between RF chokes and decoupling capacitors, or due to the inductance of capacitor leads at high frequencies. They can be eliminated by using low-Q (lossy) RF chokes, which are less likely to cause oscillations, or by using ferrite beads to add sufficient inductance to component leads or wires to dampen out unwanted VHF or UHF oscillations.

"Wide" signals are signals which are due to intermodulation distortion where the bandwidth exceeds the maximum required. The cause is usually that some amplifier stage is being overdriven, and while this may result form a design defect it is more often caused by an incorrectly adjusted microphone gain control or CW drive level. On most modern transmitters the ALC (automatic level control) voltage can be monitored on the S meter during transmit. The microphone gain or CW drive level should always be adjusted so the voltage remains within the acceptable ALC levels at all times. These levels are usually marked on the meter.

Another cause of wide signals is amateurs intentionally "opening up" the audio paths on their transmitters to allow the broadcast of wideband audio signals that exceed the 3 kHz bandwidth required for communications quality in the pursuit of "fidelity" but at the cost of causing interference to other operators.

A CW transmitter may generate key clicks if the carrier is switched on or off too rapidly when keying. The carrier should be turned on or off gently over a period of about 5 ms to avoid generating key clicks. Unfortunately even some very well regarded modern transceivers like the FT1000 MP have a problem with key clicks and may need to be modified to reduce clicks to acceptable levels.

Mains hum may be heard on transmitted signals if the power supply is inadequately filtered. The addition of a voltage regulator or additional smoothing capacitors should solve the problem.

If a transmitter is using an antenna like a long wire that is driven against earth, then it is important to have a good RF earth system that is independent of the mains earth. The mains earth wire usually travels in close proximity to the other mains wires for some distance before being physically earthed, so RF signals in the mains earth are likely to be inductively coupled to the live and neutral wires and may travel through them to neighbouring buildings, causing interference, especially to mains-operated equipment. The mains earth also often has high impedance at RF frequencies, so an independent earth system is necessary to remove RF voltages from equipment and antenna feed-lines. Of course even if you cannot provide a good RF earth, a mains ground is still required to prevent the case from having a potentially lethal voltage in the case of a fault.

Receiver Defects

The most common defect in radio and television receivers that results in interference from amateur transmissions is *receiver overload*. This is when signals stronger that the receiver was designed to handle are present at the receiver input, and inter-modulation distortion in the first mixer causes spurious products that interfere with reception.

One common cause of this is inexpensive RF masthead amplifiers that are sometimes used to improve television reception in marginal areas. While amplifiers with decent signal-handling capabilities are available, they are generally more expensive, and the inexpensive ones that are widely available are very prone to overloading.

A solution to receiver overload is to add additional filtering before the receiver that removes the strong out of band signals that are overloading the receiver. What type of filter is required will depend on what frequency transmissions are causing interference. If transmissions in the HF bands are causing the problem, then a high-pass filter between the TV antenna and the TV might solve the problem, since the TV transmissions are on higher frequencies in the VHF and UHF region, so these frequencies can be passed while blocking HF frequencies.

If amateur VHF transmissions are interfering with UHF television reception then a high-pass filter with a cutoff frequency of 470 MHz might solve the problem. However if VHF transmissions are interfering with VHF television reception, then a band-stop filter for the particular interfering amateur transmission band might be required. These band-stop filters are also called "traps". A quarter-wavelength transmission-line "stub" connected across the feed-line and open at the far end, may also serve as a trap. It presents low impedance at the frequency on which it is exactly a quarter wavelength, effectively shorting the two conductors in the feed-line together at that frequency, while presenting a high impedance at most other frequencies.

However note that if the problem is being caused by overloading a masthead RF amplifier, then no amount of filtering of the signal between the amplifier and the television will help, as in-band spurious products may already have been generated by the amplifier. In this case, replacing the amplifier with one that is more resistant to overload (or removing it altogether if reception conditions permit) may be the only option.

Interference to receivers may also result from *image signals*, also known as *second-channel* interference, if the image frequency of a receiver coincides with the frequency on which a strong amateur signal is present and the receiver has insufficient image rejection.

Common-Mode Chokes

Interference usually "gets into" the equipment being interfered with through the wires attached to it – these include antennas, speaker leads, interconnections between audio components, and mains power leads. In common–mode interference, the signal is transmitted in phase by both the conductors in the connection – for example by both the live and neutral wires in the mains, or both conductors in the speaker cable, or both the inner conductor and the earth in a coax cable.

Common-mode interference can be effectively eliminated by a common-mode choke, also known as a "braid breaker". (Although it does not involve physically breaking the braid in a coax cable, it effectively blocks the flow of common-mode signals that travel along the braid as well as in the inner conductor, which is where the name comes from).

This consists of winding several turns of the cable – which could be a mains lead, a speaker cable, or a coax cable – around a suitable core to form an inductor. Ferrite torroidial cores are the best, and are available for the purpose from local suppliers. The idea is that common-mode currents will generate a magnetic field in the core, and so the choke will act as an inductor to common-mode signals. If the inductor has sufficiently high impedance at the frequency causing the interference, then this signal can be rejected.

However differential signals – that is, signals where currents flow in opposite directions in the two conductors, for example the signal from the antenna in a TV antenna lead – will not generate a magnetic field since the fields generated by the two currents flowing in opposite directions cancel out; and so the common-mode choke does not act as an inductor for differential signals, which pass through unaffected.

Common-mode chokes can be used both with receiving equipment, such as television receivers, and with non-receiving equipment such as audio amplifiers that are suffering interference from strong radio signals.

Summary

EMC should be looked at from two perspectives: the legal (*who* is responsible for solving the problem) and the technical (*how* to solve the problem). If the interfering signal is being generated by equipment that does not need to transmit in order to function, then it is this *unintentional radiator* that is usually at fault since there are strict limits as to how much electromagnetic energy can be radiated by unintentional radiators. If the equipment being affected is not intended to receive radio signals of some kind, then it is the affected equipment that is at fault. If a signal from an intentional radiator is affecting equipment that is designed to receive radio signals, then the key question is whether the transmitter is operating within the frequency and power limits specified by its license. If the transmitter is not radiating legally, then this must be fixed. However if the transmitter is operating correctly and within license requirements, then the problem is being caused by the affected equipment responding to an out of band signal, and ultimately it is up to the owner of the affected equipment to have the problem repaired at his or her expense.

However it is advisable for an amateur whose transmissions are causing interference to assist as much as possible in diagnosing the cause of the problem and suggesting solutions. This is both to maintain a good relation with neighbours and to maintain the good image of amateur radio.

The most common transmitter problems are frequency instability, harmonic radiation, "wide" signals and key clicks. Frequency instability requires due attention in design and construction to temperature compensation, mechanical rigidity and suitable buffering of oscillators to avoid chirp. Harmonic radiation can be attenuated by a suitable low-pass filter. Wide signals are usually caused by setting the microphone gain level too high. Key clicks are the result of turning the carrier on or off too rapidly.

Receiver problems can be caused by common-mode or differential signals. Common-mode signals can be attenuated by a suitable common-mode choke (also called a "braid breaker"). Differential-mode signals require the use of suitable high-pass or band-stop filters between the antenna and the receiver. Mast-head TV amplifiers are often subject to overloading; if this occurs then the amplifier may need to be removed or replaced with one that is less subject to overloading.

Revision Questions

1 EMC defines the compatibility of electronic equipment to:
a. Static noise.
b. Man made electromagnetic noise.
c. High supply voltages.
d. Battery operated equipment.

2 The one aim of EMC is to:
a. Prevent pollution of the RF spectrum.
b. Encourage high power transmissions.
c. Discourage development of amateur radio.
d. Desensitize radio receivers.

3 Spurious oscillations caused by resonance of RF chokes can be minimized by using:
a. Low Q chokes.
b. Long power cables.
c. Non-inductive capacitors.
d. Non-resonant circuits.

4 Self oscillations can occur when the output of an amplifier is coupled to:
a. An antenna.
b. A dummy load.
c. A pi- filter network.
d. The amplifier input.

5 An RF power amplifier is found to oscillate at its fundamental frequency when the RF drive is removed. This effect is called:
a. Self-oscillation.
b. Parasitic oscillation.
c. Harmonic oscillation.
d. Overload oscillation.

6 The cure for self oscillation in an audio amplifier is:
a. To increase voltage gain.
b. To filter the feedback signal.
c. To inductively couple the input stage.
d. To introduce negative feedback.

7 Insufficient carrier suppression on an SSB Signal will cause:
a. Distortion.
b. Poor readability.
c. Difficulty to set the receiver BFO.
d. Heterodynes on the audio frequencies.

8 To minimise mains hum on transmitted signals, all DC power supplies should:
a. Use a low DC voltage.
b. Use a screened transformer.
c. Be RF decoupled.
d. Use smoothing and regulator circuits.

9 A 1 000 µF capacitor across the DC output of a power supply:
a. Will increase any 100 Hz ripple present.
b. Improve low frequency response.
c. Remove AC rectified mains hum.
d. Decrease smoothed output voltage.

10 To minimize interference on adjacent channels, voice frequencies should be kept below:
a. 500 Hz.
b. 1 kHz.
c. 3 khz.
d. 5 kHz.

11 So as not to cause unnecessary sideband splatter, the percentage modulation of an AM signal must be kept below:
a. 25%.
b. 50%.

c. 75%.
d. 100%.

12 What causes splatter?
a. Inadequate harmonic suppression in the final amplifier.
b. Excessive bandwidth of a transmitter.
c. A poorly regulated transmitter power supply.
d. Insufficient drive to the final amplifier.

13 Intermodulation caused by a linear SSB amplifier is due to:
a. Over driving the power level of the amplifier.
b. The operating frequency being to high.
c. Harmonic distortion.
d. Two modulating frequencies occurring at the same time.

14 Over-driving an SSB Linear amplifier can cause:
a. Improved communication.
b. A louder audio signal.
c. Lower power consumption.
d. Distortion and splatter.

15 Which of the following might be effective at reducing the risk of parasitic oscillations in a low power VHF output stage?
a. Ferrite beads on the emitter lead of the power device.
b. Ferrite beads on the microphone cable.
c. Ferrite beads in series with the microphone.
d. Ferrite beads on the loudspeaker leads.

16 Parasitic oscillations can cause interference, they are:
a. Of a very low frequency.
b. Always twice the operating frequency.
c. High in frequency but not related to the operating frequency.
d. Always three times the operating frequency.

17 Any non-linear device will produce:
a. Mixing products.
b. Amplification.
c. Filtering.
d. Key-clicks.

18 When a synthesized VFO oscillator is not locked to the reference frequency, it will be:
a. Stable.
b. Equal to the reference frequency.
c. Unstable.
d. Equal to the operating frequency.

19 A Domestic Receiver having an IF of 455 kHz and receiving a signal on 945 kHz, experiences strong breakthrough from someone on the 160 m band. This could be caused by second channel interference of:
a. 1,810 MHz.
b. 1,825 MHz.
c. 1,835 MHz.
d. 1,855 MHz.

20 A typical source of polluting electromagnetic interference is caused by:
a. Electric musical instruments.
b. Video signals.
c. Audio signals.
d. Arcing electrical switches.

21 A lowpass filter is most likely to be found in:
a. A crystal oscillator.
b. The output stage of an HF transmitter.
c. A TV antenna amplifier.
d. A mixer.

22 A ferrite bead around a piece of wire:
a. Decreases the wires impedance.
b. Protects the wire from damage.
c. Blocks the flow of RF signals along the wire.
d. Improves power dissipations.

23 A braid breaking toroidial choke wound onto a coax feedline:
a. Passes anti-phase currents.
b. Blocks anti-phase currents.
c. Passes in-phase common mode noise.
d. Acts as a balun.

24 An interfering signal picked up by a long feedline can be attenuated by:
a. Raising the receiving antenna.
b. Replacing the feedline.
c. Correctly matching the feedline.
d. Installing a toroidial choke.

25 In RF power amplifiers the DC wiring associated with the tank circuit often pass through ferrite beads. The beads:
a. Introduce local low pass filters in the wiring.
b. Cause high power losses at VHF.
c. Act as fine tuning controls for the tank circuit.
d. Increase the "Q" of the tank circuit.

26 To eliminate RF pickup on the outer screen of a coax cable:
a. Install a balun.
b. Remove the earth from the coax cable.
c. Install a braid breaker.
d. Use lower loss coax cable.

27 A TV antenna coax feedline picks up an amateur transmission. This can be resolved by trying to install:
a. A masthead amplifier to override the incoming interference.
b. A braid breaker.
c. New TV coax cable.
d. Filters on the mains power plugs.

28 It is found that interfering signals are being induced on the braid of an antenna downlead to a domestic FM radio by a 144 MHz transmitter. One possible solution is:
a. To fit a braid breaker filter on the antenna downlead.
b. Remove the 144 MHz transmitter earth lead.

c. To increase the 144 MHz transmitter power.
d. To fit the 144 MHz transmitter with a low pass filter.

29 The antenna of an amateur station must be located in a position that :
a. Is easily accessible.
b. Is in line with other power lines.
c. High field strengths will not be induced in domestic premises.
d. Is below all other structures.

30 The location of the feeder of an amateur antenna must be:
a. Of a precise length.
b. Kept away from other cable routes.
c. Not visible.
d. Kept close to other telephone cables.

31 The earthing of an amateur station is required to:
a. Give the mains a good earth.
b. Minimize undesired RF voltages on the feeder and equipment.
c. To prevent mains earth leakage.
d. Enable the equipment to operate from batteries.

32 When operating a mobile HF set at home from a battery supply using the base antenna, there is no breakthrough problem. When using the same arrangement with an earthed battery charger also connected, breakthrough also occurs on an electronic organ. The possible cause is:
a. The production of harmonics at the transmitter.
b. Very strong received signals.
c. Poor RF earthing.
d. RF earthing is too good.

33 To minimize harmonic radiation most HF transmitters contain:
a. A high pass filter.
b. A notch filter.
c. A low pass filter.
d. Band pass filters.

34 The term "trap" when discussing filters describes a device which:
a. Increases signal output.
b. Narrows the bandwidth of an antenna.
c. Acts as a notch filer.
d. Acts as a dummy load.

35 The length of a co-axial trap used to filter out an interfering signal is:
a. A quarter wave length of the interfering signal.
b. A random length.
c. The wave length of the transmitter signal
d. 250 mm.

36 A notch filter one quarter wavelength long used to filter out an interfering signal on the VHF bands is called:
a. A stub.
b. A balun.
c. A transformer.
d. An antenna tuning unit.

37 The main reason for providing substantial mains earthing points on radio frequency electronic equipment is:
a. To provide a path for RF to be bypassed to earth.
b. To provide a path for fault currents to be passed to earth.
c. To bypass all spurious signals to earth.
d. To increase earth resistance.

38 The leads used to connect RF equipment to earth should be:
a. Connected to the nearest mains plug earth terminal.
b. As short as possible.
c. Bare copper wire.
d. Connected via a suitable resistor.

39 In order to prevent the feeder to an antenna from radiating it should be:
a. As long as possible.
b. Cut to an exact length.
c. Screened and earthed.
d. Run close to the antenna.

40 In considering the equipment and power levels in a densely populated neighbourhood, it might be advisable to:
a. Keep the antenna as low as possible.
b. Locate the antenna as remotely as possible from the neighbours.
c. Use maximum output power.
d. Always use long feedlines.

41 The best place for an HF beam to minimize interference for an amateur living in a semi detached house is:
a. On the joint chimney stack in the centre of the roof.
b. Overhanging the next door's roof space.
c. As high and far away as possible.
d. As low and far away as possible.

Chapter 28 - Measurements

Measurements are important to determine whether equipment is operating properly and to diagnose faults. This chapter introduces some of the measurements of interest to amateurs and the test equipment we use to make these measurements.

The Ammeter

The ammeter is used to measure current. In its simplest form it consists of a coil through which the current to be measured flows, mounted on a bearing and suspended between the poles of a magnet. A current flowing through the coil will generate a magnetic field, which will interact with the magnetic field from the permanent magnet, causing the coil to pivot on its bearings. This moves a pointer attached to the coil, which indicates the current flowing on the meter scale. This is called a *moving-coil meter*.

An ammeter is connected in series with the wire in which the current to be measured is flowing, so that the current flowing through the wire also flows through the ammeter. In order to have the least effect on the circuit under test, the ammeter should have as small a resistance as possible.

The range of an ammeter can be extended by connecting a resistor, called a *shunt*, in parallel with the ammeter. The purpose of the shunt is to cause only a small part of the current being measured to flow through the meter, allowing the meter to measure a larger current than it was originally designed to. The shunt resistance can be calculated using the formula:

$$R_S = R_M / (n - 1)$$

where R_S is the shunt resistance, R_M the resistance of the ammeter, and n is the scale factor – that is, the ratio between the desired full-scale meter reading, and the full-scale reading of the meter without a shunt. For example, suppose you want to measure a current of up to 1 A using a meter with a full-scale deflection of 1 mA and an internal resistance of 100 Ω. Then the scale factor is 1 000 (to increase the full-scale deflection current from 1 mA to 1 A), so

$$\begin{aligned} R_S &= 100 / (1\,000 - 1) \\ &= 0{,}100\ \Omega \end{aligned}$$

Ammeters designed for small currents are generally called *milliammeters* or *microammeters*.

The Voltmeter

Voltmeters are used to measure voltage. A milliammeter can be converted into a voltmeter by adding a suitable *multiplier* resistor in series with the milliammeter. For example, suppose a milliammeter with a full-scale deflection of 100 μA and an internal resistance of 1 kΩ is required to read voltages up to 10 V. The total resistance of the milliammeter plus the multiplier can be found by applying Ohm's law:

$$\begin{aligned} R &= V / I \\ &= 10 / 0{,}000\ 1 \\ &= 100\ \text{k}\Omega. \end{aligned}$$

Since the internal resistance of the milliammeter is 1 kΩ, the multiplier required is 99 kΩ.

A voltmeter is used by connecting it in parallel with the component across which the voltage is to be read. In order for it to have the least effect on the circuit, the resistance of a voltmeter should be as high as possible. Transistorized voltmeters, using transistors, field effect

transistors or other devices to buffer the input can have in input resistance of many mega-ohms.

Moving coil meters are usually designed to measure DC. In order to measure AC voltages, a simple rectifier circuit may be employed. This results in the meter measuring the *average* value of the rectified AC waveform, not the RMS value. However the meter scales for AC voltmeters are usually calibrated so that if the waveform is a pure sine wave, then the scale will read the RMS value. However for waveforms other than sine waves, the reading will not be an accurate RMS value.

The Multimeter

The multimeter is a common piece of test equipment that uses a moving-coil or digital meter to measure voltage, current and resistance, usually in several ranges. Some multimeters may also measure capacitance and other quantities.

Frequency Counter

The frequency counter consists of digital circuits that count the number of cycles of the input waveform in a certain period, and then use this to calculate and display the frequency of the input signal on a digital display. The accuracy of a frequency counter depends largely on the accuracy of the internal reference oscillator used to time the counting period. If this is crystal controlled the frequency counter may be very accurate, but if it is a simple inductor/capacitor oscillator the frequency counter may have an error of several percent.

Power and SWR Meter

Power meters measure the power output of a transmitter. Depending on the meter, it may measure the *average* power or the *peak* power. The distinction is especially important for phone signals as the human voice has much higher peak amplitude than average amplitude, and this will be reflected in AM and SSB signals. (In FM signals the transmitter output is constant irrespective of the amplitude of the modulating signal.) Power meters are sometimes called *Wattmeters*.

SWR meters generally measure both forward and reflected power, and then use the ratio between forward and reflected power to calculate the standing wave ratio (SWR) which is also called the Voltage Standing Wave Ratio (VSWR). Because they measure reflected power, they are sometimes called *reflectometers*. Some modern SWR meters, called *antenna analyzers*, include a built-in low power variable frequency oscillator and a frequency counter. This makes it easy to measure the SWR at the antenna (as opposed to at the transmitter end of the feed line) and also allows measurements to be taken outside the amateur bands, as the built-in oscillator is so low-powered that it is legal for use on frequencies not allocated to amateurs.

The Oscilloscope

The oscilloscope consists of a *cathode ray tube* that displays a dot on the display. The position of the dot can be adjusted from left to right by the voltage applied to the *X deflector plates* and up and down by the voltage applied to the *Y deflector plates*. The X deflector plates are usually driven by a time-base that generates a smoothly increasing voltage, causing the dot to sweep from left to right in a period set by the user, and then to return very rapidly to the left-hand side again before starting another sweep from left to right. The Y deflector plates are driven by the input voltage, usually through an amplifier (called the *Y amplifier*), causing the dot to deflect up or down according to the input voltage.

This allows the oscilloscope to display a graph of voltage (on the Y axis) against time (on the X axis) on its screen. The time-base is synchronized by a *trigger* circuit that starts the sweep

from left to right when the input reaches a certain voltage. This synchronization means that if the input consists of a repeating waveform, then the display will "stand still" on the oscilloscope screen as each successive cycle of the input waveform traces the same pattern on the cathode ray tube display. The Y-axis is usually calibrated in volts per centimetre of deflection.

Marker Generator

A marker generator is a piece of test equipment that was used to determine the frequency of a receiver before frequency counters were available. It consists of a crystal oscillator that has been designed to generate harmonics that serve as frequency "markers" throughout the HF spectrum. For example, a marker generator might be able to generate harmonics every 1 MHz, 100 kHz or 10 kHz depending on a switch setting. The user could then find the nearest 1 MHz marker, count the number of 100 kHz markers from there, and then the number of 10 kHz markers, to get an accurate frequency reading. Almost all modern transceivers include accurate digital frequency readouts, so marker generators are rapidly becoming obsolete.

The Dip Meter

The dip meter (or grid-dip oscillator) is used to measure the resonant frequency of a tuned circuit or antenna system. It consists of a variable frequency inductor/capacitor oscillator that is laid out so that the oscillator coil is accessible (usually plugged into a socket on the outside of the dip meter) and can be brought near to the tuned circuit being tested. The frequency of the oscillator is then varied, and as the frequency approaches the resonant frequency of the tuned circuit, energy is coupled from the oscillator coil to the tuned circuit and a "dip" is noted on the meter.

The Dummy Load

A dummy load consists of a non-inductive resistor (usually 50 Ω) with sufficient power handling capability to dissipate the output of a transmitter being tested. It allows transmitter tests to be carried out without actually transmitting a signal. Transmitting a signal during testing when not strictly necessary would waste bandwidth and is poor operating practice. Be careful if you build a dummy load since most high power resistors are wire-wound. These have considerable inductance and are not suitable for RF use.

The Field Strength Meter

The field strength meter consists of a small antenna, a diode detector and a sensitive microammeter. It is used to measure the strength of (fairly strong) radio signals, for example to determine the directivity and approximate gain of an antenna. Field strength meters are generally not frequency selective and will respond to the presence of RF energy over a wide range of frequencies.

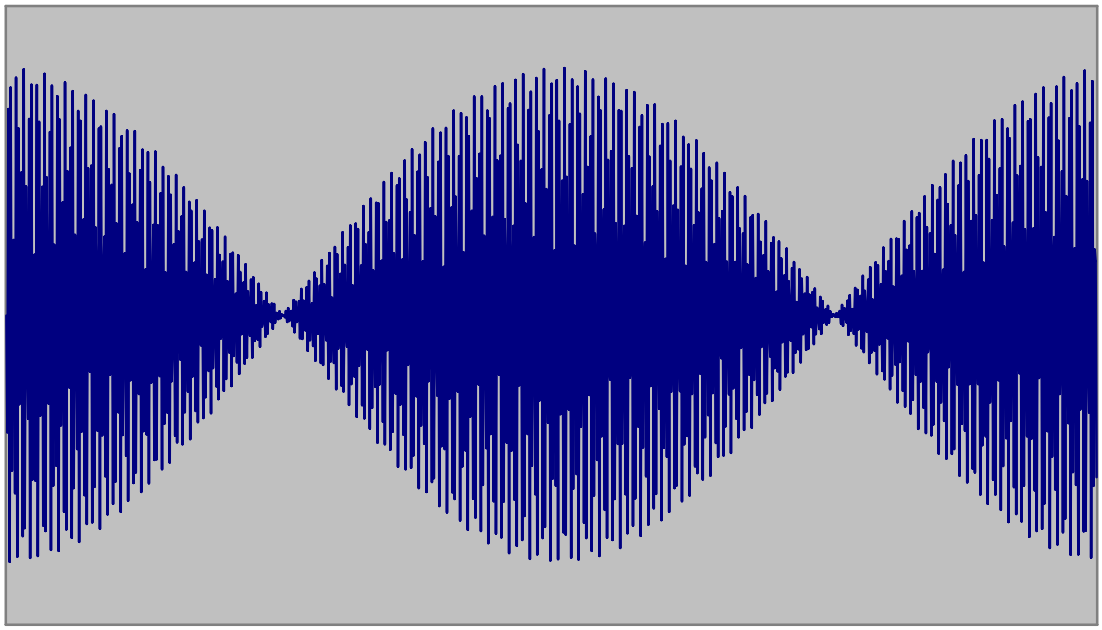
The Absorption Wavemeter

The absorption wavemeter is essentially a frequency selective field strength meter. It consists of an antenna, a tuned circuit to select the frequency, a diode detector and a microammeter. The purpose is to detect RF emissions on particular frequency bands. Because the tuned circuit is usually not very selective, it cannot be used to identify the precise frequency of a signal, but can be used to determine the approximate frequency. It is especially useful for detecting any harmonic radiation from a transmitter. For example, if you are operating a transmitter in the 80 m band but detect energy in the region of 7 MHz, then this is a good indication that your transmitter is radiating harmonics.

The Two-Tone Signal Generator

A two-tone signal generator generates an audio test signal consisting of two tones of equal amplitude that are not harmonically related. This signal is applied to the microphone input of an SSB transmitter in order to test it for linearity and to determine the peak envelope power if a peak-reading wattmeter is not available. The output of the transmitter is connected to a dummy load, and an oscilloscope is used monitor the waveform across the dummy load.

The following graph shows what the output of an SSB transmitter looks like on an oscilloscope when its input is connected to a two-tone test generator and it is operating linearly.



Output of an SSB transmitter with a two-tone test signal as input

If the output of the transmitter when viewed on an oscilloscope does not look like this, then it is not operating linearly. Specific problems include "flat-topping", when the curved tops and bottoms of the test signal are chopped off, which indicates that the transmitter is being overdriven. If successive cycles of the test signal do not join smoothly, but rather have a gap in between, then this indicates that the amplifier is incorrectly biased. Either problem will result in inter-modulation distortion and should be fixed before the transmitter is used on air.

A more accurate measure of linearity may be obtained by viewing the output of the transmitter using a *spectrum analyzer,* which breaks the signal down into its component frequencies and plots the relative amplitude of the various components against frequency.

A two-tone generator test signal and an oscilloscope can be used to measure the peak envelope power of an SSB transmission if a peak-reading wattmeter is not available. It is impossible to calculate the peak envelope power of a signal modulated by a human voice from the average power, since the peak to average power ratio differs considerably between different voices. However the peak to average ratio of the two-tone test signal is precisely 2:1. This means that if you want to set an amplifier for a maximum of 400 W PEP then you can apply a two-tone signal and adjust the amplifier until the average power output read on a wattmeter is 200 W. Then you know that the peak power is 400 W. An oscilloscope can be used to observe the amplitude of the modulation peaks. Then if voice modulation is applied,

as long as the peak output is kept at this level the peak output power (PEP) will still be 400 W.

Revision Questions

1 To extend the current range of a meter movement, a factor which must be known beforehand is the:

a. Full scale deflection voltage and coil internal resistance.
b. Maximum current-carrying capabilities of the meter movement.
c. Insulation resistance of the meter coil.
d. Maximum voltage the coil will take across its terminals.

2 To use the movement of a 0-50 microampere meter to measure voltage in the range 0 - 10 000 V, when the scale has been calibrated to read 0 - 100 V, use would be made of a:

a. Series resistor of approximately 200 MΩ.
b. Series resistor of approximately 200 000 Ω.
c. Shunt resistor of approximately 200 MΩ.
d. Shunt resistor of approximately 200 000 Ω.

3 The principal reason for using a transistorised multimeter is its greater sensitivity. On a voltage scale, this means that:

a. It will load the circuit under test to a greater extent.
b. The circuit under test sees a much higher input impedance.
c. Greater sensitivity allows the scale to be subdivided into smaller units.
d. The circuit under test will see a lower input impedance.

4 The basic instrument for measuring voltage and current is:

a. An oscilloscope.
b. A moving coil meter.
c. A field strength meter.
d. A tape measure.

5 What is a multimeter?

a. An instrument capable of reading voltage, current, and resistance.
b. An instrument capable of reading SWR and power.
c. An instrument capable of reading resistance, capacitance, and inductance.
d. An instrument capable of reading resistance and reactance.

6 How is a voltmeter typically connected to a circuit?

a. In series with the circuit.
b. In parallel with the circuit.
c. In quadrature with the circuit.
d. In phase with the circuit.

7 How can the range of an ammeter be extended?

a. By adding resistance in series with the circuit under test.
b. By adding resistance in parallel with the circuit under test.
c. By adding resistance in series with the meter.
d. By adding resistance in parallel with the meter.

8 What is a dummy load?
a. An isotropic radiator.
b. A non-radiating load for a transmitter.
c. An antenna used as a reference for gain measurements.
d. The image of an antenna, located below ground.

9 What material may a dummy load, suitable for rf, be made of?
a. A wire-wound resistor.
b. A non-inductive resistor.
c. A diode and resistor combination.
d. A coil and capacitor combination.

10 What station accessory is used in place of an antenna during transmitter tests when no signal radiation is desired?
a. A Transmatch.
b. A dummy load.
c. A low-pass filter.
d. A decoupling resistor.

11 What is the purpose of a dummy load?
a. To allow off-the-air transmitter testing.
b. To reduce output power for QRP operation.
c. To give comparative signal reports.
d. To allow Transmatch tuning without causing interference.

12 What is a marker generator?
a. A high-stability oscillator that generates a signal or series of signals from a single low-frequency signal source.
b. A low-stability oscillator that "sweeps" through a band of frequencies.
c. An oscillator often used in an aircraft to determine the craft's location relative to the inner and outer markers at airports.
d. A low-stability oscillator used for signal reception.

13 A dip oscillator is a type of:
a. RF signal generator.
b. Cathode ray oscilloscope.
c. Reflectometer.
d. RF wattmeter.

14 Which piece of test equipment contains horizontal and vertical channel amplifiers?
a. The ohmmeter.
b. The signal generator.
c. The ammeter.
d. The oscilloscope.

15 What is the best instrument for checking transmitted signal quality from a telegraphy/single-sideband transmitter?
a. A monitor oscilloscope.
b. A field strength meter.
c. A sidetone monitor.
d. A diode probe and an audio amplifier.

16 When connecting a cathode ray oscilloscope to view the wave envelope pattern of an amplitude modulated transmitter, the following coupling method would be used:

a. Direct coupling.
b. Inductive coupling.
c. Driver input coupling.
d. Inductively coupled to the final tuned circuit.

17 The vertical deflection plates in a cathode ray oscilloscope may be used to measure the amplitude of a signal. This signal displayed on the screen and measurements taken, may be calibrated and stated in terms of:

a. Current.
b. Voltage.
c. Frequency.
d. Time.

18 What kind of input signal is used to test the Peak-envelope-power of an SSB transmitter while viewing the output with an oscilloscope?

a. Normal speech.
b. An audio frequency sine wave.
c. Two audio frequency sine waves.
d. An audio frequency square wave.

19 What can be determined by making a "two-tone-test" using an oscilloscope?

a. The percent of frequency modulation.
b. The percent of carrier phase shift.
c. The frequency deviation.
d. The amplifier PEP power output.

20 What is a reflectometer used for?

a. Checking the standing-wave ratio.
b. Peaking a reciever's sensitivity.
c. Transmitter noise figure measurements.
d. Measuring sunlight intensity.

Chapter 29 - Digital Systems

Notes on using this document.

The field of amateur radio is extremely wide, and as demonstrated in this chapter, the field continues to widen. We have recognized a difficulty in trying to provide both the basic information required to pass the Radio Amateur's Examination, and providing further information to explain the more technical aspects, and to broaden learners' knowledge base. Although amateur radio is essentially a technical hobby, we should not attempt to duplicate the studies required for a proper technical diploma or degree. The RAE should be seen as an entry qualification which ensures that each entrant has sufficient knowledge to operate an amateur radio station within the relevant regulations, and to ensure that such operation is not a danger to the operator, to his equipment or to other parties.

In this chapter we have introduced a differentiation of the material into that required to pass the examination (i.e.: On which questions may be set), and further information in explanation, or of interest to advanced learners.

The main text of Chapter 29 follows the recommended syllabus for the HAREC. Questions for the Class A examination will be set only from this main text.

Additional information which may be of interest appears in blocks. Questions will generally not be set on this content.

Credits

Original text by Wessel du Preez, ZS5BLY, revised for general publication and use by RAE lecturers and candidates in July 2007 by Peter Hers, ZS6PHD.

Input is gratefully acknowledged from Colin de Villiers, ZS6COL, Ean Retief, ZS1PR, George Honiball, ZS6NE, Mark Zank, ZS6YES, Marten du Preez, ZS6ZY, Mickey Esterhuysen, ZS5QB, and Rassie Erasmus, ZS1YT.

The author wishes to thank Steve Ford, WB8IMY, Production and Editorial Manager of the ARRL for permission to use diagrams and text portions from their publications.

1. Principles of Digital Signal Processing

Signal processing is the action of modifying or enhancing one or more parameters of a signal to improve and select a wanted parameter. In radio engineering this may entail any of the familiar operations such as modulation, filtering, mixing, detection, etc. All these functions may also be accomplished digitally with the aid of a computer. Although common microcomputers may be used for this purpose, specialized digital signal processing micros have been developed that execute the required operations much faster.

Sampling

Before any signal processing can take place, the analogue signal has to be converted to a digital signal. This is done by taking periodic samples of the analogue signal and storing the instantaneous values as digital numbers. The process of sampling is illustrated in figure 29-1 below.

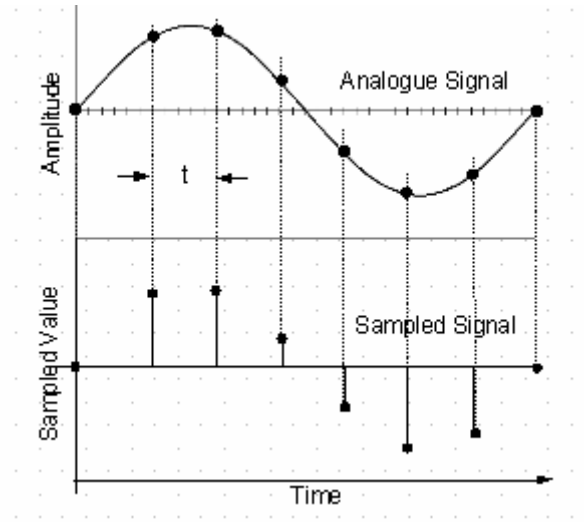


Fig. 29-1 Sampling process

Note that the sampling period "t " is much shorter than the period of the sampled wave – eight samples have been taken during this single cycle. Connecting the upper points of the samples produces a fair representation of the original at the same frequency.

In the general case the sampling rate is governed by the *Nyquist Sampling Theorem.* This requires that at least two samples be taken during each cycle in order to faithfully reproduce the original signal.

What happens if we take less than two samples during a single cycle?

The reconstructed wave is still a sine wave, but of a different and lower frequency and does not represent the original signal. It is known as an *alias frequency.* In some applications deliberate use is made of this lower frequency and it is then called *under sampling.*

In the case of a complex signal, it means that the sampling rate must be more than twice the highest frequency contained in the signal.

In order to avoid aliasing, a low pass or band pass filter is inserted ahead of the sampling device. A practical example of this is the sound card in a computer. To be able to adequately cover the audio frequency range of 20 Hz to 20 kHz, the sampling frequency is 42 kHz.

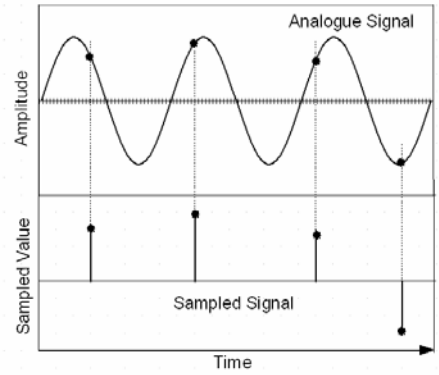


Fig 29-2 Effect of under-sampling

Analogue to Digital Conversion

The device used to convert the analogue signal to a sampled digital version is an a*nalogue-to-digital converter (ADC)*. For each sample, the ADC produces a digital number that is directly proportional to the amplitude of the input voltage. The number of bits available in the binary word limits the number of discrete voltage levels that can be resolved. An 8-bit ADC can only resolve 256 levels (2^8) while a 12-bit unit can resolve 4 096 (2^{12}) levels. This limits the resolution of the ADC as it can only report the analogue value to the nearest discrete level.

The difference between the actual and reported value is called the *quantitization error* and for a good ADC is ± ½ the value of the Least Significant Bit (LSB) for that converter. This error is pseudo-random and appears in the output as *quantitization noise.*

Once the signal has been digitized, a digital process can be implemented to perform various functions amongst which are modulators, mixers, AGC systems and filters.

The reporting process of an analogue to digital converter is illustrated in the diagram of figure 29-3 below:

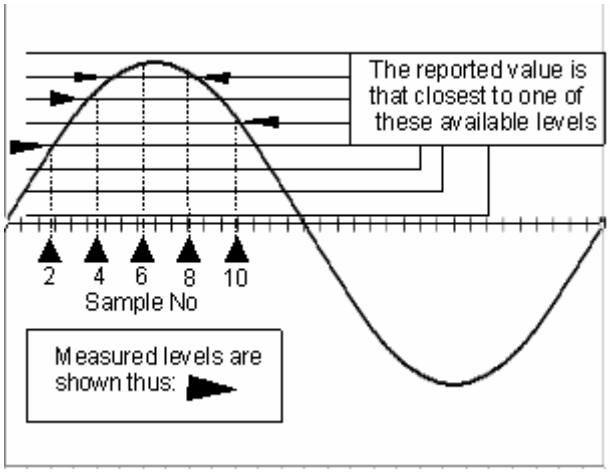


Fig. 29-3 A/D converter

A further source of noise, especially in VHF signals, is due to *aperture jitter* which is caused by small variations in the sampling clock intervals. It is, however, much smaller than the quantitization noise. Yet another source of error is caused by the non-linearity of the conversion and this is typically ±1 or 2 bits over the entire range.

Digital to Analogue Conversion

After the digital processing has been completed, the resulting digital string of words must be converted back to an analogue value. This is accomplished by a *digital-to-analogue converter (DAC)*. The device accepts a digital word as input and, at the application of a clock signal, outputs the corresponding analogue value. The discrete nature of the digital input results in a stepped analogue output that may be smoothed by filtering.

Digital Filters

Most of the following information has been gleaned from *"Experimental Methods in RF Design"* by Wes Hayward, W7ZOI, Rick Campbell, KK7B, and Bob Larkin, W7PUA. The book has been published by the ARRL and the material is reproduced here with the kind permission of the ARRL who holds the copyright.

As mentioned previously, DSP can be used to implement a number of analogue functions. Figures 29-4 and 29-5 below show the alternate implementations of a band-pass filter.

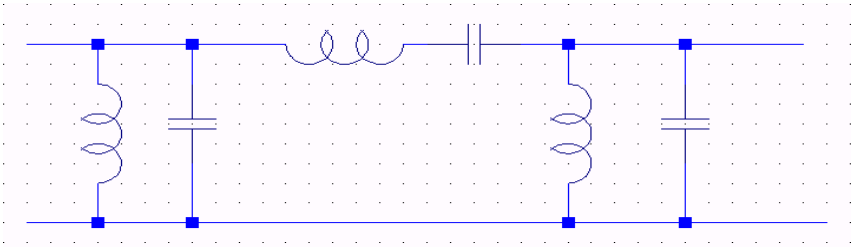


Fig. 29-4 Analogue implementation of a band-pass filter

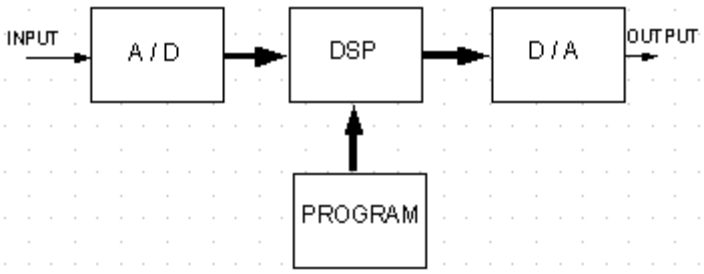


Fig. 29-5 DSP implementation of a band-pass filter

Digital filters are implemented in one of two ways, called IIR and FIR filters. The difference between them is that IIR filters involve results of previous calculations (feedback) while FIR (*Finite Impulse Response*) filters do not. IIR stands for *Infinite Impulse Response* which refers to the response of the filter to an impulse consisting of a single unity amplitude sample pulse. The name is somewhat of a misnomer as it would indicate that the output of the filter will last forever while it actually gets smaller until it falls below the resolution of the processor. The pulse also equates the filter response to its frequency response via the Fourier transform. In the discussions that follow, use will be made of the following symbols in order to explain the operation of digital filters.

Multiplication of Inputs

Summation of Inputs

One Sample Period Delay

IIR Filters

The simplest IIR filter is the analogue of the R-C low-pass filter shown below. The block diagram shows the simple IIR filter that has the same response as the R-C filter.

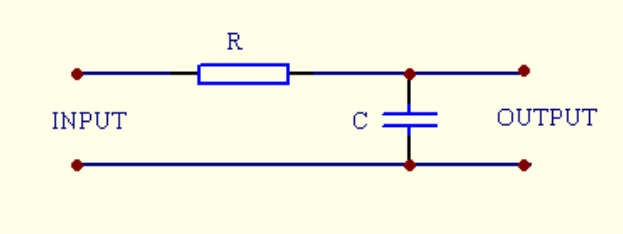


Fig. 29-6 Analogue low-pass filter

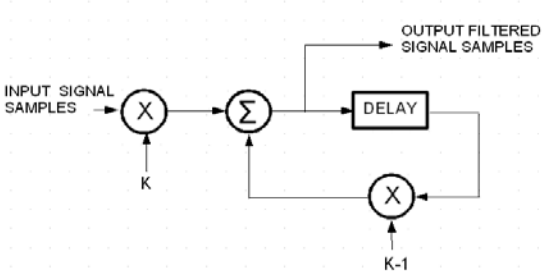


Fig. 29-7 Simple IIR low-pass filter

If we call the digital input sample x_i and the filter output y_i, then our filter consists of the single calculation:

$$y_i = K\,x_i + (1-K)\,y_{i-1}$$

where K is a constant between 0 and 1 but typically 0,001 or less.

Demonstrating how it works

Using the above equation we can now calculate the operation of this filter for the first few terms as the input rises from 0 to 1. Assume that the output is 0 when we start and choose K = 0,1 to make things happen faster.

New Input, x_i	K x_i	(1-K) y_{i-1}	New Output,
0,0	0,0	0,0	0,0
1,0	0,1	0,0	0,1
1,0	0,1	0,09	0,19
1,0	0,1	0,171	0,271
1,0	0,1	0,244	0,344

The calculation shows that the output is growing towards 1 but with a smaller step after each new input. This is the same exponential growth as that of an R-C circuit. Figure 29-8 below shows the charging characteristics of the R-C filter as well as that of the digital circuit.

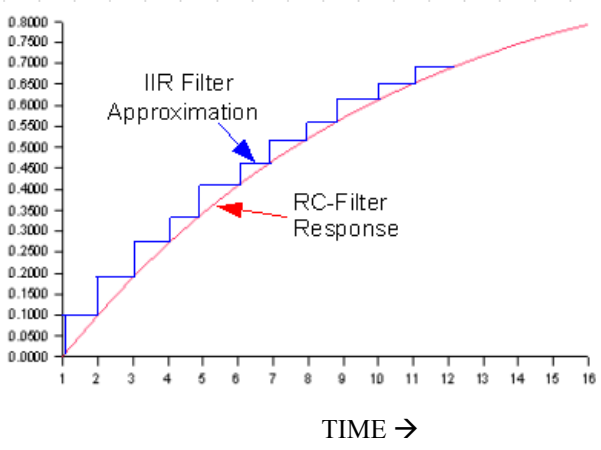


Fig. 29-8 Response of analogue and DSP filter

FIR Filters

For more complex filters it is often desirable to use the FIR filter, standing for *Finite Impulse Response.* These filters don't use the previous outputs of the filter computation, but do use the current input along with many of the previous inputs.

DSP implementation of the FIR filter is very simple as shown in the block diagram of figure 29-9 below:

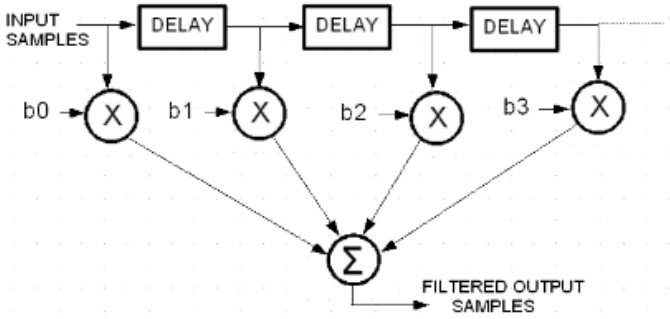


Fig. 29-9 FIR Filter

The input signal is available in digital format from the A/D converter. A delay line consists of places in memory for some number of previous samples. Each time a new sample arrives, it is placed in the beginning of the delay-line memory. Multiplying all the samples by constant numbers and then adding them together forms new outputs. The constant multiplier numbers (b1, b2, b3, etc.) are referred to as the FIR coefficients, or tap weights. The filter design consists of choosing these coefficients to suit the particular application. As with analogue filters there are trade offs between the number of coefficients, pass-band ripple and the out-of-band rejection.

The FIR structure can be used to form filters that are highly selective to the frequency of a sine wave input signal. All of the response characteristics of L-C filters, such as Butterworth, Chebyshev and others are possible with the FIR filter.

Direct Digital Synthesis

Direct Digital Synthesis (DDS) is the process of generating arbitrary waveforms by means of digital methods. The word "direct" refers to the fact that no feedback is used in the basic process. It is based on the fact that, for a sine wave, there exists a constant relationship between the phase and the amplitude values of the wave.

We first consider the analogue view. Visualize a point on the circumference of a rotating wheel. Each rotation of the wheel represents one cycle of the waveform and the speed of rotation corresponds to the frequency (or cycles per second, Hz). If the wheel rotates at a constant speed, then a sine wave will be produced. The basic operation is shown in figure 29-10 below. In this diagram the horizontal axis represents the phase or angular displacement of the circle radius. The amplitude above (or below) the zero line is transferred to the graph as well. The relationship between the angle Ø, the radius of the circle, r, and the amplitude a is given by:

$$\text{Sin } Ø = a / r$$

The amplitude is then given by $a = r \text{ Sin } Ø$ and hence the term "sine wave".

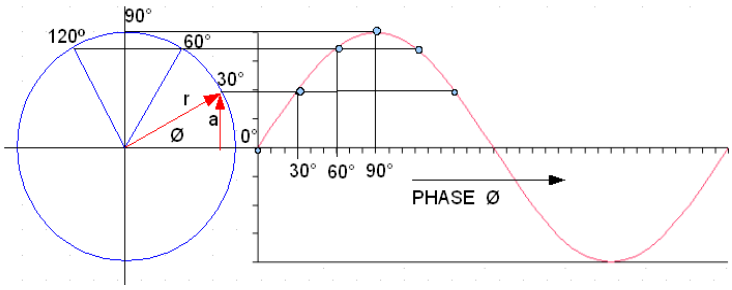


Fig. 29-10 Generation of a sine wave

Note that the amplitude values start repeating after 90° eg. the amplitude at 120° is the same as that for 60°. After rotating through half of a revolution, the values again repeat but are inverted.

Creating a Sine Wave by Direct Digital Synthesis - DDS

Figure 29-11 shows a DDS system to generate a sine wave at a given frequency by digital integration of a phase increment.

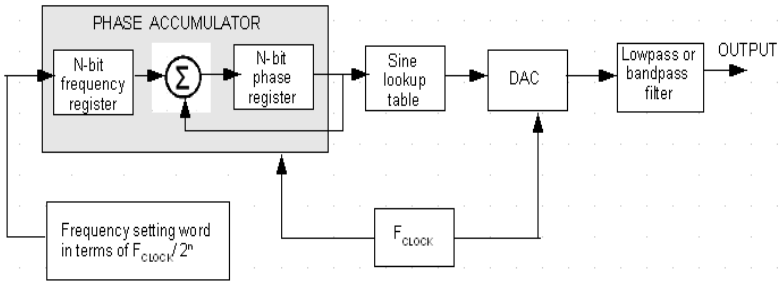


Fig. 29-11 Block diagram of a DDS system

The integration is not a continuous process as would be performed by op-amps, but is stepwise. This phase increment acts as a frequency setting word since it determines the number of clock cycles per complete cycle of the wanted frequency. The phase accumulator accepts an N-bit binary frequency setting word and outputs a binary word describing the instantaneous phase to the sine lookup table which in turn outputs the corresponding amplitude value for that phase location to the digital to analogue converter. The phase accumulator is incremented with every clock pulse and, at the same time it transfers the digital amplitude value to the DAC. The output of the DAC is a stepwise sine wave which is filtered by the output filter. If N bits are available to the phase accumulator then the output is given by:

$$F_{OUT} = (\text{Frequency Setting Word}) \times F_{CLOCK} / 2^N$$

The lowest possible output frequency as well as the frequency increment is $F_{CLOCK} / 2^N$ which occurs when the FSW is 1. Using a clock frequency of 100 MHz and N= 32 gives rise to a frequency increment of 0, 0232 Hz !

As the clock frequency is usually derived from a stable crystal oscillator, the DDS can deliver output frequencies with a very high stability and resolution. The maximum output frequency is usually limited to about half of the clock frequency and sometimes even lower. The stepwise output also generates unwanted signals of which most may be eliminated by the judicious choice of the clock frequency and output filters.

Figure 29-12 below shows the spectrum of the stepwise output before filtering.

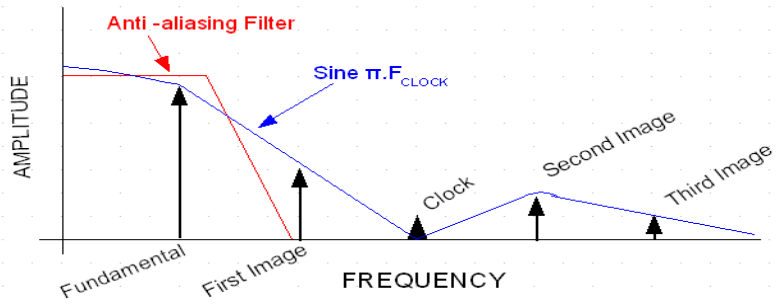


Fig. 29-12 DDS OUTPUT SPECTRUM

The Fourier Transform

A Fourier transform is a mathematical technique for determining the frequency content of a signal. It was originally developed by Joseph Fourier (1768-1830) for continuous signals. For sampled signals, such as used in DSP, a variant of the transform called the *discrete Fourier transform* (DFT) is used. Basically the DFT converts a block of N input samples into a block of N output bins.

As the transform makes use of complex sinusoids, it is very demanding on computational processing. Using the direct method of computation requires N complex multiplications and additions per bin and for N bins it means that the computation effort is proportional to N^2. After realising that the complex sinusoid was periodic with N, mathematicians (notably Carl Runge 1856-1927) further simplified the method and reduced the computational effort. This method lay dormant until it was revived by Cooley and Tukey during the 1960's. It is known as the *Fast Fourier Transform or FFT* and is in common use on many microcomputers.

Discontinuities and Windowing

When using a system of blocks of data for analysis, discontinuities exist at both the beginning and end of the blocks which cause unexpected spectral components to appear in the output. One way of alleviating this problem is to increase the number of bins, but this unfortunately complicates the decision as to which bin the result is to be allocated.

To minimise that problem use is made of a technique called *windowing*. The data block is multiplied by a *window function* which removes the sharp transitions in the envelope after which it is processed in the normal way. The best known and used window functions are the *Triangular, Blackman, Hamming, Hanning and the Rectangular*. Using a rectangular window is the same as not using a window at all while all the others have a shaping effect. The spectra of these windows all resemble that of a low-pass filter and are often used to design such filters. The diagram below shows the shape of the Hanning window.

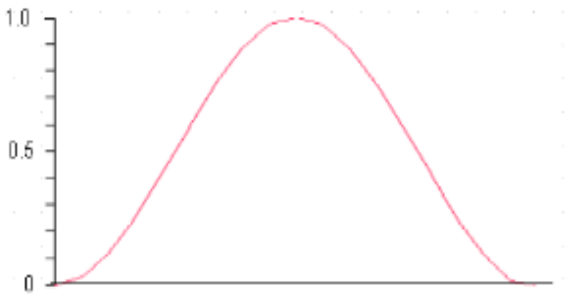


Fig. 29-13 Hanning (or Cosine Bell or Cos²) window

2. Digital Communication Modes

In addition to the normal voice communication modes, several digital modes are also in use in radio communication. These can broadly be divided into two categories namely *text modes* and *image modes*. In each of these categories several modes have been developed that are software controlled and make use of the sound card in a PC.

The input to the sound card is coupled to the audio output of the transceiver via the loudspeaker or any other dedicated audio output connector. Audio from the sound card output is fed to the microphone or other modulation input of the radio. Provision is also made to key the transmitter from the serial port of the computer. Algorithms within the software control the signal processor in the sound card to manipulate the audio frequencies and/or phase of the incoming and outgoing signals.

The main analogue modulation methods are covered in Chapter 20 and will not be repeated here.

A. Text Modes

Morse Telegraphy

This is the original mode for both amateur and commercial radio communication and operates by on-off keying of the transmitter carrier wave. It occupies a very narrow bandwidth and is readable even under very marginal conditions. The signals are human readable between 5 and 60 words per minute.

Radio teletype (RTTY)

This is one of the first data communication modes that received widespread use. The five bit code could only encode $2^5 = 32$ symbols which could not deal with 26 letters, 10 figures and the punctuation marks. The problem was solved by using one or more of the codes to select symbols from code-translation tables and so produce upper-case letters, figures, special symbols and punctuation marks. The coding used by most amateur stations is the *International Telegraph Alphabet No 2 (ITA2)* also known as *Baudot.*

Due to the mechanical nature of the early machines the signalling rates are not very high and vary between 65 and 133 words per minute. It is common to measure the signalling rate of RTTY in *"baud"*. The signalling rate in baud is the reciprocal of the shortest pulse length. Figure 29-14 below shows the formation of the letter "Y ". Typically, the pulse length is 22 ms which equates to 45,45 baud, a common signalling rate. In Europe a signalling rate of 50 baud was common.

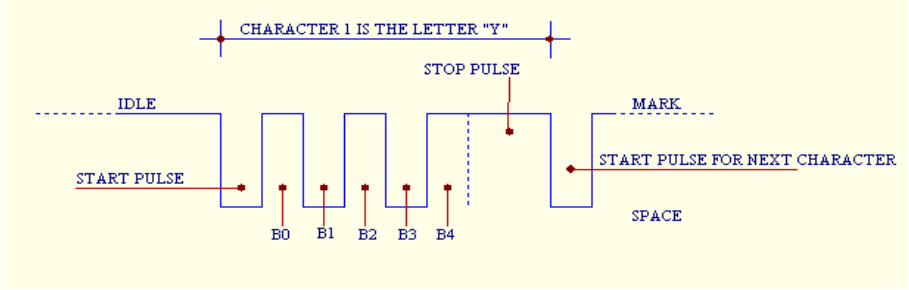


Fig 29-14 Character "Y" in Baudot code (ITA2)

Early systems used *Frequency-shift Keying* (FSK), moving the carrier frequency by 850 Hz. Modern systems employ *Audio Frequency-shift Keying* (AFSK) using an audio tone of 2 125 Hz for a mark level and 2 295 Hz for the space level. It is used on VHF as well as on HF SSB frequencies.

AMTOR

Amtor is based on a system devised in the Maritime Mobile Service to improve communications between stations using RTTY. This was necessary in order to overcome the problems experienced by RTTY when signal fading and noise occurred on the channel. The system converts the 5-bit code to a 7-bit code in such a manner that there are four mark and three space bits in every character. There are two modes that are commonly used in AMTOR. *Mode A* uses automatic repeat request in which the receiving station acknowledges the received characters by checking for the correct 4/3 bit ratio or else calls for a repeat. *Mode B* uses a simple forward error correction by sending each character twice.

ASCII

This is the *American Standard Code for Information Interchange* which is commonly used for information-processing systems (computers), communication systems and related equipment. ASCII uses a seven bit code that can accommodate $2^7 = 128$ symbols. Although it is not strictly a part of the ASCII standard, an eighth bit may be added as a parity (error checking) bit. By using other methods for error checking, the eighth bit may be used to extend the set of ASCII symbols to 256.

In radio communication the ASCII character set is mostly used for serial data transmissions. These transmissions may be either synchronous or asynchronous. For synchronous transmission the character code is sent on its own while for asynchronous transmission a start and one or two stop bits is added. The standards for serial transmissions require that the character is transmitted with the least significant bit first and that the start and stop pulse duration is the same as that of the information pulse.

The signalling rates differ depending on the medium used but in amateur radio the following are most widely used: 75, 110, 150, 300, 600, 1 200, 2 400, 4 800, 9 600, 16 000, 19 200 and 56 000 bits per second. Signalling speed is often quoted in baud which is equal to one discrete condition or event per second.

Some digital modulation methods have more than the normal two states. In the dibit modulation method two ASCII bits are sampled at a time and have values of 00, 01, 10 and 11. The four phase modulation method assigns phase of 0º, 90º, 180º and 270º respectively. For this type of phase modulation the signalling speed in bauds is half the transfer rate in bits per second. Many modern modulation techniques have been developed that exploit this multi-bit encoding.

ASCII code example.

As an example the code for " M " is 1001101. A pulse sequence showing the serial transmission of the letter "M " using ASCII code is shown in figure 29-15 at right.

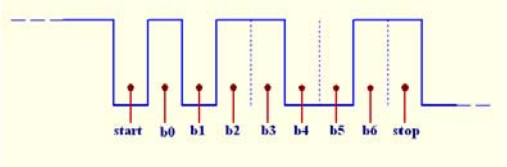


Fig 29-15 ASCII character "M " Setup = 7-bit, 1 stop, no parity

PACKET RADIO

Data communications is telecommunications between computers. *Packet switching* is a form of data communications that transfers data by subdividing it into "packets", and *packet radio* is packet switching using radio. (Definition by Steve Ford, WB8IMY)

This tells us that amateur packet radio is the communication between computers using amateur radio stations and that the computer operators are radio amateurs. The system uses only a single channel to service multiple communications simultaneously and is readily connected into networks that cover a given area. At the hub (or node) of each area we find a packet bulletin-board system (PBBS) which can store data and also forward it to other hubs. A possible configuration is shown below.

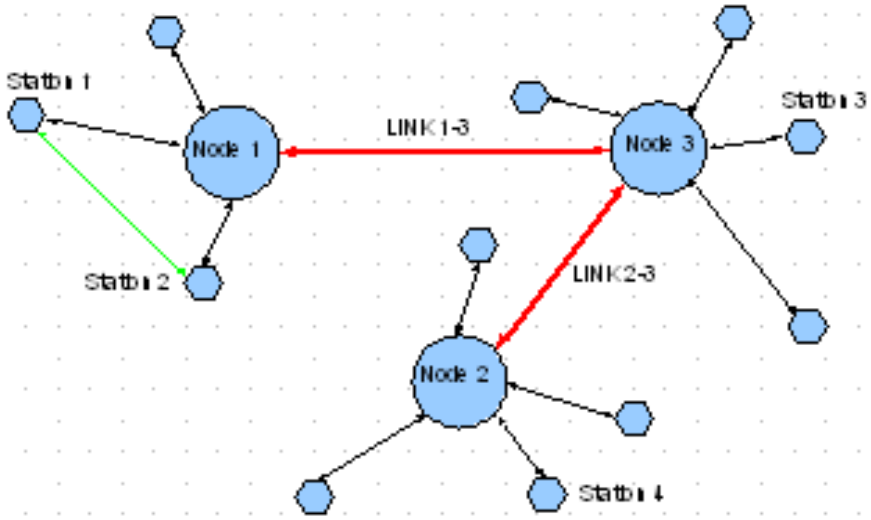


Fig. 29-16 Layout of a Packet Network

Stations may communicate directly or via one or more nodes. Nodes can be connected in national as well as international network configurations. In order to maintain compatibility and also ensure virtually error free data communications, packet radio uses a modified C.C.I.T.T. (International Telephone and Telegraph Consultative Committee) recommendation X-25 protocol called AX-25. The main difference is that the address frame in AX-25 can accommodate amateur call signs and has an added unnumbered information (UI) frame. This protocol specifies the format of a packet radio frame and the action a station must take when it transmits or receives such a frame.

FLAG	*ADDRESS*	*CONTROL*	*INFO/MSG*	*FCS*	*FLAG*
0111111	14 to 70 bytes	1 byte	Message or data up to 256 characters	Frame Check Sequence 2 bytes	0111111

Destination	Source	Digipeaters, Nodes, Paths
ZS6XXX-9	ZS5XYZ	

The heart of any packet radio system is the *Terminal Node Controller* or TNC. The function of the TNC is to take the asynchronous data from the computer or terminal (usually in ASCII form) and assemble it into packets or frames. These frames are then passed on to a modem for conversion into audio tones which in turn are fed to the radio transmitter. During reception the reverse takes place.

It is not always necessary to use a TNC as several computer programmes have been developed that use only a simple modem while the assembly and disassembly of the frames are done by the computer.

Most TNC's can also be used to re-transmit the received packets to other stations; this is called *digipeating*. These stations do not add any information and merely re-transmit any frame that contains their call sign in the *digipeat* portion of the address field in the frame. In order to handle network operation, the native firmware in the TNC-2 is replaced with firmware called NET/ROM. This firmware supports the network and transport layers (levels 3 and 4) of the packet-radio network.

Packet bulletin boards, normally referred to as a BBS, have facilities to store messages which may be retrieved at a later stage by the addressee or, if placed in a public area, by any station requesting such a file. Examples of the latter are newsletters or other shared information.

One of the most popular applications of packet radio is TCP/IP which stands for *Transmission Control Protocol / Internet Protocol.* It is actually a set of several protocols that provide a flexible and adaptable means of networking. The *Telnet* protocol allows for a chat session, while the *File Transfer Protocol (FTP)* allows the transfer of files between stations. A large number of software sets for TCP/IP is available based on the original NOSNET by Phil Karn, KA9Q .

APRS

The APRS or *Automatic Position Reporting System* was developed by Bob Bruninga, WB4APR, to overcome the limitations of packet radio when used for real time communications. The limitations are mainly that a permanent link between all stations would be required for packet radio. APRS uses the UI frames so that any number of stations may participate in the exchange of information.

For many events the position of the radio is important and provision has been made to add a GPS (*Global Positioning System*) receiver to the system enabling moving targets to be tracked. This is extremely useful for search and rescue operations as control stations are able to track the position of rescue workers and vehicles on a computer map in the control center. Some systems also allow the data from digital weather stations to be added to the transmitted data.

PSK31

PSK is a realtime keyboard-to-keyboard mode of operation. It was developed by Peter Martinez, G3PLX, and derives its name from the fact that it uses phase shift keying at a baud rate of 31,25 baud. A new variable length code was developed for PSK in which the most used characters have shorter codes. The mode is very robust and can maintain communication even under very adverse propagation conditions.

Error correction was added to PSK by using quaternary *phase shift keying* (QPSK) and a convolutional encoder to generate one of four different phase shifts that represent five successive data bits. A Viterbi decoder is used at the receiving end to correct errors. This decoder tracks the 32 possible (5 bit) sequences, retaining only the most likely ones while discarding the other.

WSJT

WSJT is a weak signal communications program. It supports FSK441, a fast digital mode for meteor scatter as well as JT44, a slow digital mode for troposcatter and earth-moon-earth (EME) communication. An EME echo mode is included for measuring your echoes from the moon. The programme was written by Joseph H Taylor Jr., K1JT. The ability of these programmes to extract data almost below the noise level have made them popular with experimenters and have enabled both long distance VHF and UHF as well as moon bounce using modest station equipment.

CLOVER

Clover is an advanced HF digital system developed by Ray Petit, W7GHM. It uses a four-tone modulation scheme and allows different modulation formats to be selected manually or automatically depending on the prevailing signal conditions. The block data rate is about 250 bits per second.

PACTOR

PACTOR combines some of the best features of Packet Radio and AMTOR, with some further additions.

Some benefits of PACTOR include: Operation at 100 or 200 baud, depending on path conditions; a 16-bit cyclic redundancy check to provide near error-free operation; memory ARQ in which the controller is able to combine parts of successive blocks to eliminate errors; and selective use of IRA data compression.

PACTOR occupies the same bandwidth as 300 baud HF Packet.

Originally copyrighted by SCS GmbH, the mode was released into the public domain in 1991. Subsequent versions –II and –III with increased speed and other capabilities remain proprietary.

You can listen to samples of various digital modes including Pactor©-I, -II and –III at http://www.wb8nut.com/digital.html.

B. Image Modes

Facsimile

This is the name for methods that are used to transmit very high resolution still pictures using voice bandwidth radio channels. It is the oldest of the image transmitting methods and has been the main method used to transmit weather charts and newspaper photos by radio. It is also used by polar orbiting weather satellites to transmit their ground and cloud images to earth. FAX transmissions are made up of 800 to 1 600 scanning lines, which provide a higher resolution than the home TV.

Modern amateur radio applications are based on software utilising the sound card in a PC which then displays the image on the computer monitor. The high resolution is achieved by slowing down the data rate resulting in transmission times of 4 to 10 minutes per image.

Slow-scan Television

As the name implies, this mode differs from the normal commercial television with respect to the scanning rate of the picture frames. Home TV in South Africa (PAL) produces 25 complete images consisting of 625 lines per second and occupies a bandwidth in excess of 6 MHz.

In order to transmit an image, very accurate and low distortion audio tones in the range of 1 500 Hz to 2 300 Hz are used to represent the picture element (pixel) intensities and are changed at the correct pixel rate. A 1 200 Hz tone is used for synchronization. These audio tones are then used via audio frequency shift keying (AFSK) to modulate a transmitter. On the receiving side the audio tones are decoded by measuring their frequency. Most modern systems make use of the computer sound card and appropriate software to code and decode the various modes that are in use at present. (at least 12 exist at the time of writing!)

In order to utilise normal voice transmission bandwidth, early experimenters devised an 8 second transmission standard. Audio tones in the range 1 500 to 2 300 Hz were used to represent the range of shades between white and black. A 5 ms pulse of 1 200 Hz indicated the start of a line and a longer pulse at the same frequency announced the start of a new frame, which was made up of 120 lines.

The same standard was transferred to colour images by placing red, green and blue filters in front of the camera and sending each colour image separately and assembling them at the receiving end. It was known as the *frame-sequential* method. Any interference during the transmission of any frame could cause the entire image to be useless and an improved method of *line sequential* transmission was adopted. In this case each line was sent three times, each time using a different filter. The basic waveform for these methods is shown in figure 29-17 below.

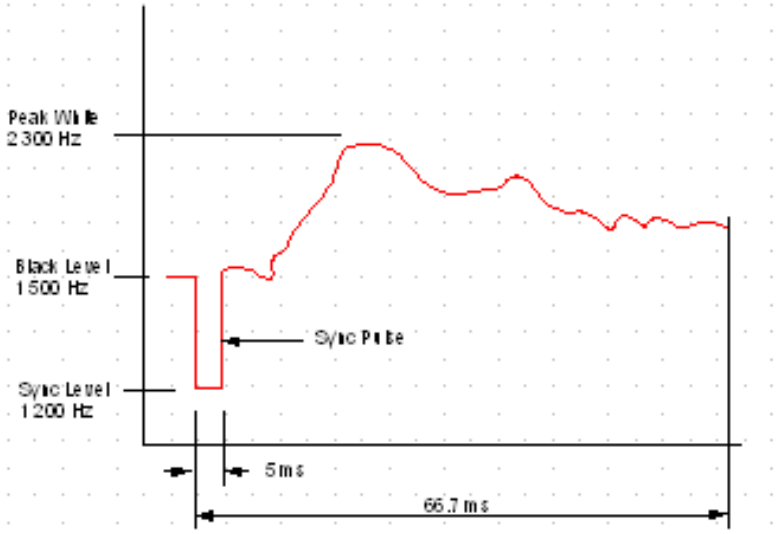


Fig. 29-17 Monochrome SSTV signal

Later developments used luminance and chrominance signals instead of the usual red-green-blue signals. The first 50 to 70% of the scan line contains the luminance information which is a weighted average of the RGB signals. The remaining part of the line contains the chrominance or colour information. This made the method compatible with the monochrome system as it could use the first part of the line to display the monochrome image and ignore the chrominance information. Although this encoding method reduced the time to transmit a frame from 24 s to 12 s, it produces bad quality when it encounters sharp, high contrast edges. The newer modes have therefore returned to RGB encoding.

Fast-Scan Television.

Fast-scan television in the amateur service is a wide-band mode that follows standard broadcast scan rates.
Due to the wide bandwidth, this mode is restricted to the UHF and microwave bands.

C. Digital Voice – The Future!

Digital modes offer advantages over their analogue counterparts and are widely used in most modern voice communication systems such as the public switched telephone network (PSTN) and cellular networks. Digital detectors need only to ascertain whether the received signal represents a digital zero or a digital one. Coding schemes have also been devised to detect and correct possible errors in the transmission making the method robust even under poor propagation conditions. The digital signals are also readily processed by advanced methods using DSP techniques to enhance their quality.

Digital speech coding can be classified into two types namely *waveform coding* and *source model coding*. Waveform coding techniques involve the quantitization of the speech waveform at high rates to produce high quality speech at the cost of a high bit rate. An 8-bit A/D converter sampling at 8 000 samples per second would produce a bit rate of 64 000 bits per second! More complex voice coders or *vocoders* use Adaptive Differential Pulse Code Modulation (ADPCM) in which prediction with differential quantitization is used to reduce the data rate to 24 – 32 000 b/s.

Model coding vocoders use a parametric model to approximate short segments of speech (between 10 and 40 ms). The speech is modelled with a fundamental frequency, a set of

spectral coefficients and a set of frequency dependant voicing decisions. This multi-band voicing information and algorithms to analyse and synthesize speech has resulted in the availability of Advanced Multi-band Excitation (AMBE) vocoders that can provide high quality speech at rates between 2 000 and 5 000 bits per second. These low bit rates allow the digital signal to be transmitted by HF radio whilst not exceeding the normal 2,5 to 3,0 kHz bandwidth.

Such a system was developed for amateur radio application in 1998 by Charles Brain, G4GUO, and Andy Talbot, G4JNT. Their system is based on an AMBE1000+ speech vocoder operating at 2 400 b/s plus 1 200 b/s of forward error correction. Thirty-six tone carriers spaced at 62,5 Hz are used producing an overall bandwidth of 312,5 Hz to 2 500 Hz using DQPSK modulation.

A typical application using a vocoder as in the above system is shown in Fig 29-18 below.

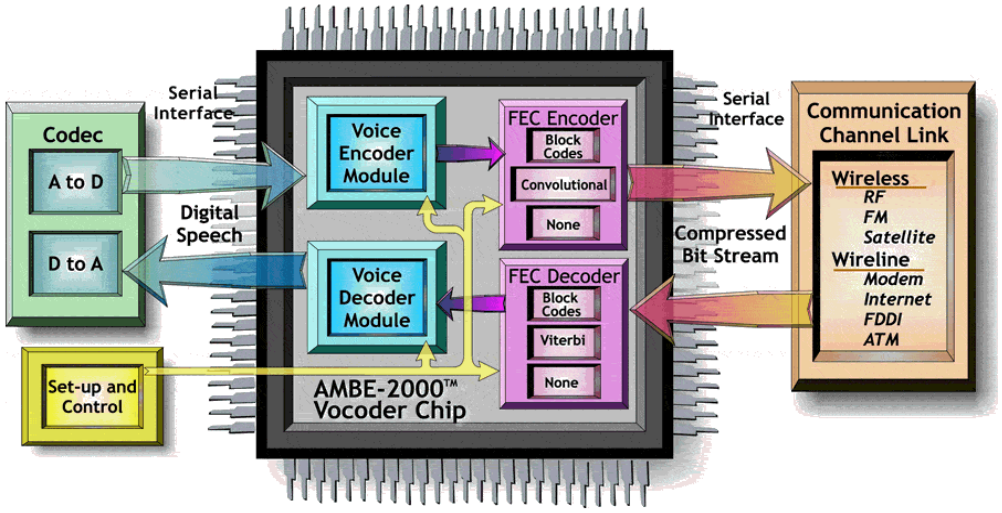


Fig. 29-18

A major advantage of a digital voice system is that it may be added to any SSB transceiver (even the older valve types) without the need for any internal modifications to the radio. In most cases the use of digital voice can provide better signals than the addition of a linear amplifier.

The use of digital voice in amateur radio is still in its infancy and ideally lends itself to experimentation and development.

VoIP – Voice over Internet Protocol

Voice over Internet Protocol (VoIP) is not a radio communication mode or modulation type, and is mentioned here only as it is utilized in various networks which combine radio communication with communication via the internet.

VoIP, also called IP Telephony, Internet telephony, Broadband telephony, Broadband Phone and Voice over Broadband is the routing of voice conversations over the Internet or through any other IP-based network.

Amateur radio has adopted VoIP by linking repeaters and users with Echolink, IRLP, D-STAR, Dingotel and EQSO. Echolink and IRLP are programs/systems based upon the Speak Freely VoIP open source software. Echolink allows users to connect to repeaters via their

computer (over the Internet) rather than by using a radio. By using VoIP Amateur Radio operators are able to create large repeater networks with repeaters all over the world where operators can access the system with ham radios.

References and Recommended further reading

1. Hayward,W., Campbell,R. and Larkin,B., *Experimental Methods in RF Design,* ARRL, Newington, CT, 2003.

2. Reed, D.G. (Ed), *The ARRL Handbook for Radio Communication,* ARRL, Newington, CT, 2005.

3. Smith, D., *Digital Signal Processing,* ARRL, Newington, CT, 2003.

4. Higgins,R.J., *Digital Signal Processing in VLSI,* Prentice Hall, Englewood Cliffs, NJ. 1990.

5. Oppenheim, A.V. And Schafer, R.W., *Digital Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1975.

6. Shanmugam, K.S., *Digital and Analog Communication Systems,* John Wiley & Sons, NY, 1985.

7. Struik, D.J., *A Concise History of Mathematics,* Dover Publications Inc., NY, 1994.

8. Fletcher, W.I., *Engineering Approach to Digital Design,* Prentice Hall, Englewood Cliffs, NJ, 1980.

9. Smillie, G., *Analogue and Digital Communication Techniques,* Arnold, London, 1999.

10. Ford, Steve. *HF Digital Handbook,* ARRL, Newington, CT, 2004.

11. Taggart, Ralph E, *Image Communications Handbook,* ARRL, Newington, CT, 2002.

12. *AMBE-2000 Vocoder Chip User Manua*l, Digital Voice Systems Inc., Westford, MA, 2005.

Appendix 29A - Number Systems and Logic Operations

In understanding digital systems it is essential to understand the difference between a *digital* and an *analogue* signal. The world around us is mostly an analogue world. When the sun rises in the morning, the light intensity varies continuously from dark to light with time. This is an analogue signal as it is *continuously changing with time*. On the other hand, operating the light switch in a dark room will result in an instantaneous change from dark to light. There are thus only two states, dark and light. This is called a digital signal because it is *discontinuous with time*.

The light-dark digital system illustrated above is called a *binary* system as at any given time the light intensity can only assume one of two states: light or dark. Most electronic systems use the binary system because many physical systems are readily described by two state levels EG: switches that are on or off, a mark or no mark on a paper, Etc. By using some form of coding, these two states can be used to represent any number.

Number Systems

Any number system has two distinct characteristics: a set of *symbols* (digits or numerals) and a *base* or *radix.* A number is a collection of these digits and the value of the number is a weighted sum of the digits. The weight of a digit is determined by the base or radix of the system and the position of the digit with respect to the separator (decimal point or comma).

Our decimal system, with which we are all familiar, is a base-10 system with ten symbols: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. To count , we start at 0 and work our way up to the highest symbol, 9. To represent the next number, after adding another one to this 9, we have to resort to our position value or radix. In the decimal system the columns to the left of the separator have weights of 1, 10, 100, etc. and those to the right of the separator have weights of 1/10, 1/100, 1/1000, etc. In scientific notation we write:

1 000 = 10^3
100 = 10^2
10 = 10^1
1 = 10^0
(Separator)
1/10 = 10^{-1}
1/100 = 10^{-2}
1/100 0 = 10^{-3}
Etc

Thus the value of the number 438 is calculated as follows:

$$4 \times (10^2) + 3 \times (10^1) + 8 (10^0) = 4 \times 100 + 3 \times 10 + 8 \times 1 = 438$$

In the binary system we only have two symbols for which we traditionally use 0 and 1. These symbols are known as *bits* which is a contraction of the words **Bi**nary digi**t** With a radix or base of two, our columns now have the following weights **if we read from right to left:**

2^0 = 1
2^1 = 2
2^2 = 4
2^3 = 8
2^4 = 16

The decimal value of the binary number 1011 is calculated as follows:

$$1 \times (8) + 0 \times (4) + 1 \times (2) + 1 \times (1) = 8 + 0 + 2 + 1 = 11$$

As in the decimal system, the left-most digit is called the *most significant bit, MSB* and the right-most bit is called the *least significant bit or LSB*. When using binary systems, four bits are called a *nibble* and two nibbles or eight bits are called a *byte.* A *word* may consist of two or more bytes.

Large numbers in binary tend to get rather long and other systems have been created to make the numbers more human friendly and easier to work with. A system that is used extensively in digital systems, such as computers , is the *hexadecimal* system. As the name implies, it is a base-16 system and requires 16 symbols. Our decimal system only has ten symbols so we add alphabetic letters to obtain the other. The binary, decimal and hexadecimal values are then as follows:

BINARY	DECIMAL	HEXADECIMAL
0000	0	0
0001	1	1
0010	2	2
0011	3	3
0100	4	4
0101	5	5
0110	6	6
0111	7	7
1000	8	8
1001	9	9
1010	10	A
1011	11	B
1100	12	C
1101	13	D
1110	14	E
1111	15	F

When using the hexadecimal notation, it is usual to denote it by following the number by an "H" as in 3FA4 H or a subscript as in $3FA4_H$.

Our familiarity with the decimal system and the need to have binary numbers for our computers have led scientists to devise another number system which could ease the entering of digital numbers into a machine. The most widely used system is the ***B**inary **C**oded **D**ecimal* system or BCD for short. In this system each decimal digit is expressed as a 4-bit binary number therefore the decimal numbers 0 to 9 are coded as 0000 to 1001. Note that this incomplete binary coding causes the BCD notation to lose the mathematical relationship of the weighted sum. Direct binary mathematical operations are therefore not possible with BCD numbers and specialized circuits are used for this purpose

Physical Representation of Binary States.

In electronic systems, state levels are physically represented by voltages. A typical choice for a system operating on 5 V would be:

State 0 = 0 V
State 1 = 5 V

It is, however, unrealistic to obtain these precise voltages so a more practical choice is a range of values such that:

State 0 = 0, 0 to 0,5 V
State 1 = 2,4 to 5,0 V

with the level between 0,5 and 2,4 V being undefined.

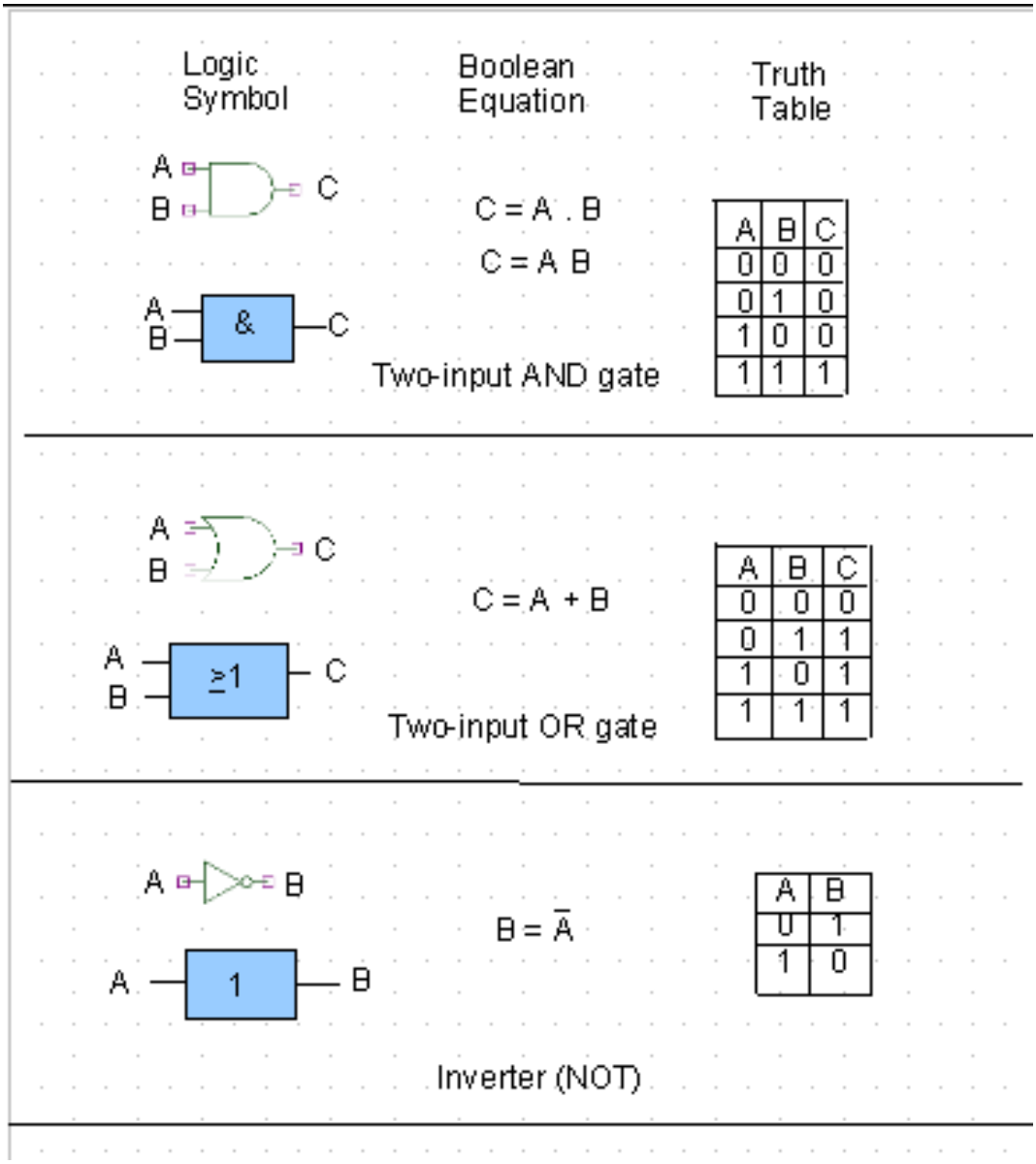
Over the years a number of digital or logic families have been developed and they differ in their defined levels of an 0 or 1 as well as their switching speeds, their drive capabilities and power consumption. Prevalent today is the **C**omplementary **M**etal **O**xide **S**emiconductor or CMOS family. These devices operate over a wide range of supply voltages and usually define the state levels as a percentage of the supply voltage. They are also very fast in operation.

Logic Operations

Digital circuits combine binary inputs to produce a required binary output or combination of outputs. These range from simple combinations of 0's and 1's to sophisticated computations as used in computers. The circuits fall into two distinct categories namely *combinational* or *sequential* types. In *combinationa*l circuits the output depends only on the present state of the inputs while in *sequential* logic the outputs depend on the present inputs, the previous sequence of inputs and often also on a timing or clock input.

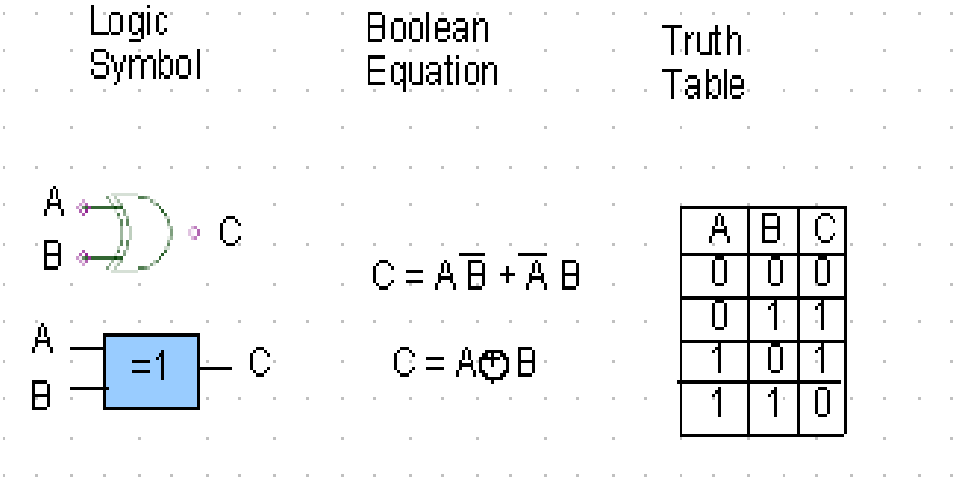
1. Combinational Logic

Combinational circuits are composed of logic gates which perform binary operations. The design of these circuits are based on *Boolean algebra,* named after George Boole, who developed the system. Just as normal algebra has a set of basic operations such as addition, subtraction, multiplication and division so Boolean algebra has its own set of logical operations. They are AND, OR and NOT and may be described by either a Boolean equation or a truth table. The basic Boolean or logic operators are shown in the figure below. The logic symbols are shown in both their conventional (top symbol) as well as the IEEE symbol (bottom symbol) for each operator together with the Boolean equation and a truth table.



Boolean operators

Looking at the truth table for the AND function we note that the output C will be high (or true**) if, and only if,** both A and B are high (true). In the case of the OR function the output is true if **either** A or B or both A and B are true. We will now add a further function which is called an E*xclusive OR*. The corresponding diagrams and truth table are shown in the figure below:



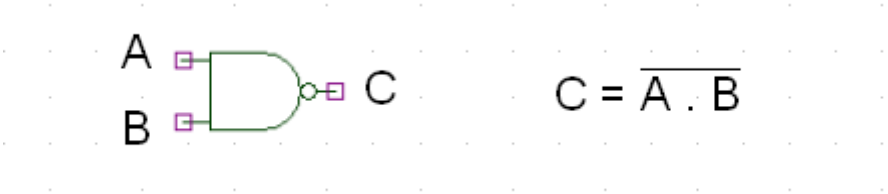
XOR (Exclusive OR)

In the above, note that the output, C, is only true if **either** A or B is true and **excludes** the case where both inputs are true.

A further refinement in logic gates is to add an inverter after the standard function to create inverted outputs. The table below shows these functions:

AND + INVERTER = NAND
OR + INVERTER = NOR

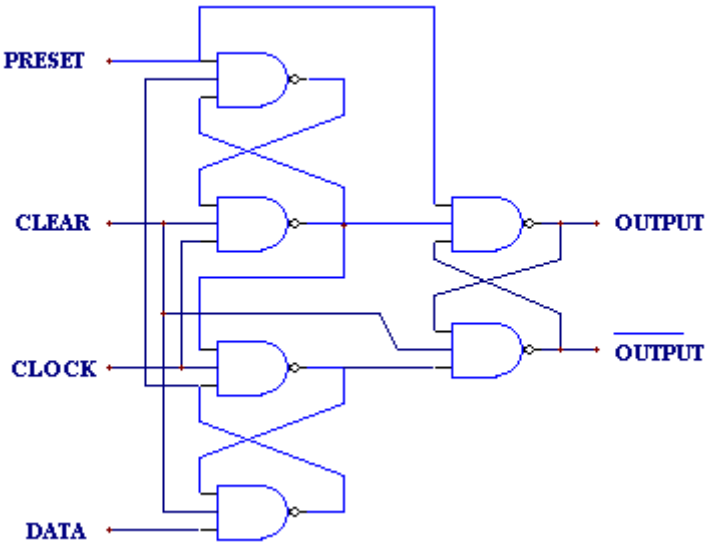
The symbols for these functions are the same as the base function with the circle of the NOT function attached to the output port and for the NAND function is as shown in the figure below with the Boolean equation. The line above the inputs is used to indicate the inverted input and is read as "NOT A and B".



The NOT function

2. Sequential Logic

In sequential logic circuits the outputs depend on the present inputs, the previous sequence of inputs, as well as a timing or clock signal. The basic building block of these circuits is the *flip-flop* of which there are a number of variations. Our main building block is the D-type flip-flop. This is a circuit built up of gates and has two inputs and (usually) two outputs. An example of such a flip-flop is shown in the figure below and, in this case, is constructed using NAND gates.

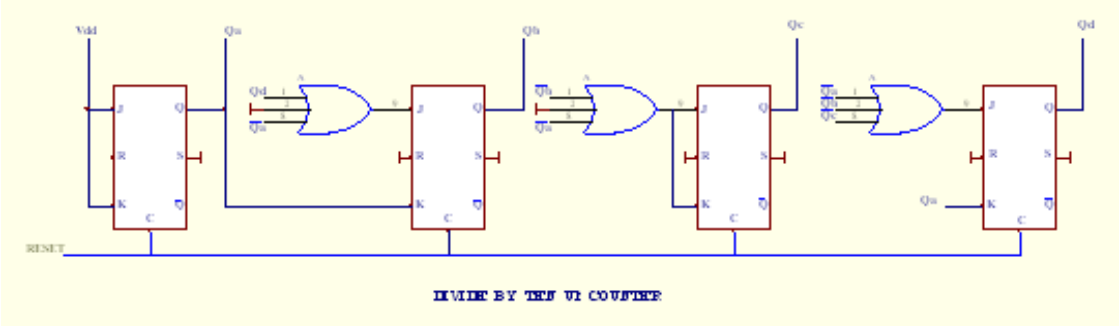


D-Type Flip-flop

At switch-on, the state of the output is undetermined so a positive pulse is applied to the **Clear** input. This sets the output to a 0. If a clock pulse is now applied, the rising edge of the clock will transfer the level present at the **Data** input to the output. Note that the two outputs are complements of each other.

By connecting such, or other more complex flip-flops together and gating their outputs, complex logic sub-systems may be constructed. These include counters, shift registers, arithmetic logic units, memories and, ultimately, complete microcomputers.

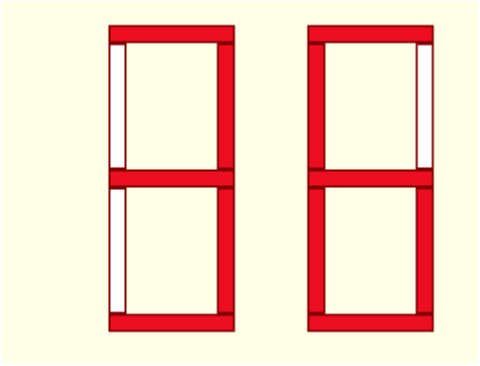
For illustrative purposes only, the figure below shows the construction of a "divide-by-ten" up-counter together with the corresponding truth table. Note that the outputs, Qa to Qd, represent the number of clock pulses received in binary coded decimal format. The counters may be cascaded to accommodate more than one decimal figure.



State	Qa	Qb	Qc	Qd
1	0	0	0	0
2	1	0	0	0
3	0	1	0	0
4	1	1	0	0
5	0	0	1	0
6	1	0	1	0
7	0	1	1	0
8	1	1	1	0
9	0	0	0	1
10	1	0	0	1

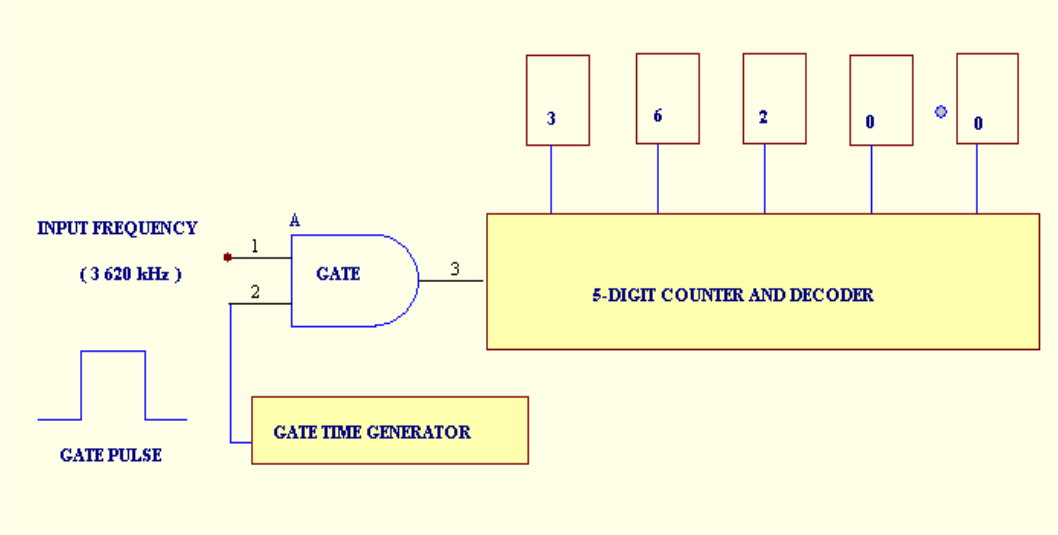
Divide by 10 up-counter and truth table

By decoding the outputs of the counter to illuminate one or more of the elements in a seven segment display unit, the counter contents become readable by human operators. A readout using this kind of display is shown below.



7-segment Display showing the number 36

It now becomes an easy matter to construct a device which can display the frequency of a signal by counting the number of clock pulses that enter the counter in a given time period. In the diagram shown below we apply a frequency of 3 620 kHz to the gate at the input to the counter. If the gate pulse is set to open the gate for 10 ms, then during that period, the counter will receive 36 200 pulses and display the count as indicated. By changing the gate pulse time, the number of pulses counted and displayed may be varied to suite the application. The position of the decimal point is usually controlled by the gate time generator selector and often also indicates the range of the counter, Eg: kHz as in our example.



5-digit Counter

Appendix 29B - Standard ASCII Codes

The table below shows the 7-bit ASCII character code set. Symbols shown as NUL to US are the so-called control characters that are often used to control devices such as tape recorders, printers, etc.

			B6 →	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**
			B5 →	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**1**
			B4 →	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
B3 ↓	**B2 ↓**	**B1 ↓**	**B0 ↓**								
0	**0**	**0**	**0**	**NUL**	**DLE**	**SP**	**0**	**@**	**P**	**'**	**p**
0	**0**	**0**	**1**	**SOH**	**DC1**	**!**	**1**	**A**	**Q**	**a**	**q**
0	**0**	**1**	**0**	**STX**	**DC2**	**“**	**2**	**B**	**R**	**b**	**r**
0	**0**	**1**	**1**	**ETX**	**DC3**	**#**	**3**	**C**	**S**	**c**	**s**
0	**1**	**0**	**0**	**EOT**	**DC4**	**$**	**4**	**D**	**T**	**d**	**t**
0	**1**	**0**	**1**	**ENQ**	**NAK**	**%**	**5**	**E**	**U**	**e**	**u**
0	**1**	**1**	**0**	**ACK**	**SYN**	**&**	**6**	**F**	**V**	**f**	**v**
0	**1**	**1**	**1**	**BEL**	**ETB**	**`**	**7**	**G**	**W**	**g**	**w**
1	**0**	**0**	**0**	**BS**	**CAN**	**(**	**8**	**H**	**X**	**h**	**x**
1	**0**	**0**	**1**	**HT**	**EM**	**)**	**9**	**I**	**Y**	**i**	**y**
1	**0**	**1**	**0**	**LF**	**SUB**	*****	**:**	**J**	**Z**	**j**	**z**
1	**0**	**1**	**1**	**VT**	**ESC**	**+**	**;**	**K**	**{**	**k**	**{**
1	**1**	**0**	**0**	**FF**	**FS**	**,**	**<**	**L**	****	**l**	**|**
1	**1**	**0**	**1**	**CR**	**GS**	**-**	**=**	**M**	**]**	**m**	**}**
1	**1**	**1**	**0**	**SO**	**RS**	**.**	**>**	**N**	**^**	**n**	**~**
1	**1**	**1**	**1**	**SI**	**US**	**/**	**?**	**O**	**_**	**o**	**DEL**

7-bit ASCII character code set

Revision Questions

Note: The answers to these questions are not in multiple choice format. The intention is to give you an idea of the type of questions that may be asked from this chapter. In the exam the answers will be presented in multiple choice format.

1. Describe how an analogue signal is converted into a digital signal.
2. How is an appropriate sample rate determined?
3. What is the relationship between the number of bits for an ADC and the number of discrete levels which can be resolved?
4. What is quantitization error?
5. If you have a 12-bit ADC, what level of quantitization noise can be expected?
6. What device converts a digital signal into an analogue signal?
7. Give the names of two ways in which digital filters may be implemented.
8. Can a low-pass filter be implemented as a digital filter?
9. Can a high-pass filter be implemented as a digital filter?
10. The output of a digital-to-analogue converter will contain noise due to the finite steps in the analogue output. How can this noise be removed?
11. Describe the process of direct digital synthesis.
12. Can you generate a sawtooth wave by direct digital synthesis?
13. Mention three digital communication modes used to transmit text.
14. What character codes are normally used in digital communication?
15. Describe packet radio.
16. What protocol is used in packet radio?
17. List some advantages of digital voice transmission.
18. Describe the operation of IRLP or Echolink.
19. Why do you need to be a licenced radio amateur to use IRLP?
20. Why do you need to be a licenced radio amateur to use Echolink?

Chapter 30 - Operating Procedures

This chapter introduces the basic operating procedures used in the amateur bands. After completing the chapter you should understand the procedures used to communicate with other amateurs on HF phone, CW and on VHF repeaters.

Safety Considerations

Mains voltages can kill. So when you install equipment, ensure that:

- Insulated wire with a suitable voltage rating is used for all connections.
- All exposed metal surfaces (including equipment cases) are properly earthed.
- When mains power is switched, a two-pole switch is used to switch both the live and neutral lines.
- The plugs used have a suitable current rating for the equipment, and that mains outlets are not overloaded by having too many plugs connected.

You should have a single "master switch" that can be used to turn off the mains supply to all your equipment and that is known to everyone who lives with you. This will allow your family to safely disconnect the mains supply in the event that you are incapacitated by an electric shock.

Much of the older equipment, and most modern linear amplifiers, use valves. These circuits typically have potentially lethal high tension (HT) voltages of between 500 and 3 000 V. So do not poke around inside valve equipment unless the power is off, the equipment is disconnected, and you are sure that the capacitors in the power supply are discharged. (The power supply should have *bleeder resistors* connected across each of the capacitors to discharge them, but these may be missing or faulty).

The RF output of a radio can give a very nasty RF burn. In fact, radio-frequency signals are used by some medical instruments to cut human flesh. So always turn your radio off before adjusting antennas, or loading coils, or doing anything else that may lead you into contact with the RF output.

HF Phone Procedures

The Phonetic Alphabet

The phonetic alphabet is used whenever information must be spelt out. It should be used for callsigns when initiating a contact. Once it is clear that the other station has got your callsign correct then you can revert to normal pronunciation ("ZS1AN" instead of the phonetic "Zulu Sierra One Alpha November".

The standard phonetic alphabet is:

Alpha
Bravo
Charlie
Delta
Echo
Foxtrot
Golf
Hotel
India
Juliet
Kilo

Lima
Mike
November
Oscar
Papa
Quebec
Romeo
Sierra
Tango
Uniform
Victor
Whisky (or *Water* in Muslim countries)
X-Ray
Yankee
Zulu

These words have been chosen so they are easily distinguishable from one another. This, plus knowing the words that are included in the phonetic alphabet, aids intelligibility in poor conditions. For example, if you hear only "elta" then you know the word must have been "Delta". These advantages are lost if non-standard phonetics are used, so you should always use the standard phonetic alphabet.

Initiating Contacts

Before calling, you should listen for at least 30 seconds to see whether the frequency is clear. If you do not hear anyone else on or near the frequency, then you can ask whether the frequency is clear:

> *Is this frequency in use? Zulu Sierra One Alpha November.*

Wait another thirty seconds and if you have not heard anything then you can proceed to call "CQ" to ask for a contact.

> *CQ CQ CQ this is Zulu Sierra One Alpha November, Zulu Sierra One Alpha November standing by.*

Wait for 5 seconds. If you do not receive a response, then call again. If after you have tried many times you still have not received a response, then this may indicate that propagation conditions are poor on the band you have chosen.

If you want you can make a *directional* call, which means asking for only certain stations to reply. If you call *CQ DX* this means you are asking for only "long distance" (DX) contacts, which usually means stations on another continent. If you call "CQ Europe" or "CQ Germany" then you are asking only for stations from a particular continent or country to reply.

Replying to a CQ

If you hear a station calling CQ and you would like to make contact, then before you call check the following:

1. Is it a directional call, and if so are you in the right area to respond? For example, a South African station should not respond to "CQ Japan" but may respond to "CQ Africa" or to "CQ DX" from a non-African station.

2. Make sure you know where the station is listening for a response. Usually this will be on the same frequency as their call. However rare DX stations may work "split"

which means they are listening on a different frequency, generally higher than the one they are calling on. For example, if you hear a DX station call "Sierra Tango Zero Romeo Yankee up five" it means the operator will be listening 5 kHz higher than the frequency they called on. You will need to know how to activate the "split" function on your transceiver to work this station.

3. Ensure that a suitable antenna is connected and (if necessary) that your antenna tuning unit (ATU) is correctly set for the frequency and antenna. If the ATU is not set then *do not* tune up on the frequency where you hear the CQ call, as this is most inconsiderate and will cause interference to the station calling. Rather change frequency by at least 3 kHz to an unoccupied frequency, and after checking that the frequency is not in use, tune up there and then return to the frequency where you heard the call.

Of course it is wise to check these things *before* you search for stations calling CQ, so when you hear one you can respond immediately. Suppose you hear W1XX calling and having checked everything you are ready to call. Then you would say:

Whisky One X-ray X-ray this is Zulu Sierra One Alpha November, Zulu Sierra One Alpha November, Zulu Sierra One Alpha November standing by.

Note that the callsign of the station being called is always given *first*, and the callsign of the station calling comes *second*. This is important and getting it wrong will mark you as a poor operator.

I think it is unnecessary to repeat the callsign of the station you are calling several times – after all, she or he presumably knows his/her callsign, and unless conditions are bad, once is normally sufficient for them to hear that you have got it right. Under poor conditions, however, you might want to repeat it a couple of times.

Exchanging Reports

After making contacts, the first things stations do is usually to exchange signal reports and basic information such as the name and location of the operator. Signal reports are exchanged according to the standard Readability-Strength (Tone) code, usually abbreviated RST. The Tone part is only used for CW communication, so for Phone it is RS – Readability and Strength only. The meaning of the RST values are as shown below:

READABILTY
1 -- Unreadable
2 -- Barely readable, occasional words distinguishable
3 -- Readable with considerable difficulty
4 -- Readable with practically no difficulty
5 -- Perfectly readable

SIGNAL STRENGTH
1 -- Faint signals, barely perceptible
2 -- Very weak signals
3 -- Weak signals
4 -- Fair signals
5 -- Fairly good signals
6 -- Good signals
7 -- Moderately strong signals
8 -- Strong signals
9 -- Extremely strong signals

TONE
1 -- Sixty cycle a.c. or less, very rough and broad
2 -- Very rough a.c. , very harsh and broad
3 -- Rough a.c. tone, rectified but not filtered
4 -- Rough note, some trace of filtering
5 -- Filtered rectified a.c. but strongly ripple-modulated
6 -- Filtered tone, definite trace of ripple modulation
7 -- Near pure tone, trace of ripple modulation
8 -- Near perfect tone, slight trace of modulation
9 -- Perfect tone, no trace of ripple or modulation of any kind

The Q-code "QTH" is often used to mean the location of the operator, although this is actually incorrect usage as the Q code should only be used in Morse code, and normal plain language should be used in Phone. So you might hear the following reply from W1XX:

> *Zulu Sierra One Alpha November this is Whisky One X-ray X-ray. Thanks for the call, you are five and six, fifty-six. My name is Bob, Bravo Oscar November, and my QTH is Boston, Massachusetts. ZS1AN from W1XX.*

The signal report indicates that our signal is perfectly readable, with good signal strength. You would reply with a signal report, and also your name and location.

> *W1XX from ZS1AN Good morning Bob, thanks for the report. Your RST is five nine, five nine here in Cape Town. My name is Andrew, Alpha November Delta Romeo Echo Whiskey, Andrew. I'm testing a new rig here, a Kenwood TS850S, running 100 W into a triband Yagi at 15 metres. Back to you Bob. W1XX this is ZS1AN.*

And so the conversation continues. You must by law identify your station on each separate transmission (each "over"). However once you are sure that the other station has got your callsign correct you can use plain language instead of the phonetic alphabet.

Ending the QSO

"QSO" is also from the Q code, and it means a contact between two stations, which may include several transmissions ("overs") by each station. The end of the conversation will probably go something like this:

> *W1XX from ZS1AN. Well Bob many thanks for the nice chat, I must be off now. I will QSL via the bureau. 73 to you and your family and see you later. W1XX this is ZS1AN clear but listening for your final.*

> *ZS1AN from W1XX. Fine business Andrew, nice to meet you and enjoy that new radio, it sure sounds good from here. 73 until next time from W1XX signing clear.*

Here "73" is from an old telegraph code meaning "best wishes". (Note that it is already plural, so you should *not* say "73s"). An alternative when addressing someone of the opposite sex is "88", which means "love and kisses". "QSL via the bureau" means send a QSL card confirming the QSO via the QSL bureau. The QSL bureau is a delivery service for QSL cards that operates via the amateur radio societies in many countries, including the South African Radio League in South Africa. Again it is generally poor procedure to use abbreviations designed for Morse code when operating phone, but "QSL" and "73" are universally used, so you will just have to accept them as an exception to the rule!

After the QSO
If you have not already filled in your log during the QSO, then you should do so immediately after the QSO. Remember that you are required by law to keep a log of all HF transmissions (including unanswered CQs, but no-one actually logs these).

If you have offered to send a QSL card, then it is a good idea to write it out immediately, as it can become a chore if you wait for hundreds of QSOs to accumulate before writing out the cards.

What Not to Do

Don't put out endless streams of "CQ" calls. It is most irritating to hear

> *CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ This is Zulu Sierra One Alpha November, Zulu Sierra one Alpha November, CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ This is Zulu Sierra One Alpha November, Zulu Sierra one Alpha November, CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ CQ This is Zulu Sierra One Alpha November, Zulu Sierra one Alpha November, CQ CQ...*

It is acceptable to repeat a CQ call twice if you must. We don't really see the purpose. We would rather make a short CQ, then leave a gap for a response, and call CQ again if We do not receive a reply.

Don't say "Over" or (worse) "Over and out". That really pegs you as a beginner.

Never ever say "good buddy" or "10-4" or any other CB jargon!

CW Procedures

CW QSOs generally follow a similar format to HF phone, but with many more abbreviations! There are two main kinds of abbreviations used: the Q Code, which uses three-letter groups starting with the letter Q to stand for questions and answers; and informal abbreviations for commonly used words. Let's start with the Q Code. Each entry can be used either as a question – in which case it is followed by a question mark – or as a statement, which may be in response to the question. For example, "QTH?" means "what is your location?" and the reply might be "QTH Cape Town" meaning "my location is Cape Town". You should know the following abbreviations:

Q-Code	Question	Statement
QRG	What is my exact frequency?	Your exact frequency is...
QRL	Are you busy?	I am busy.
QRM	Are you being interfered with?	I am being interfered with.
QRN	Are you troubled by static?	I am troubled by static
QRO	Should I increase power?	Increase power.
QRP	Shall I decrease power?	Decrease power.
QRQ	Shall I send faster?	Send faster.
QRS	Shall I send more slowly?	Send more slowly.
QRT	Shall I stop sending?	Stop sending.
QRU	Have you anything for me?	I have nothing for you.
QRV	Are you ready?	I am ready.
QRX	When will you call me again?	I will call you again at .... hours
QRZ	Who is calling me?	You are being called by ....
QSB	Are my signals fading?	Your signals are fading.
QSL	Can you acknowledge receipt?	I acknowledge receipt.

QSP	Will you relay (a message) to ... ?	I will relay to ...
QSY	Shall I change frequency to ... kHz ?	Change frequency to ... kHz.
QTH	What is your location?	My location is ...

Note that QRX also has the informal meaning “standby”. The difference between QRM and QRN is that QRM means “man-made interference”, while QRN means “noise”. A “QRP” station means a station transmitting with low power, usually 5 W or less.

Initiating Contacts

First listen for at least 30 seconds to see whether the frequency is in use. If nothing is heard, then ask whether the frequency is in use by sending:

QRL? DE ZS1AN

“QRL?” means “are you busy?” or “is this frequency in use?”. “de” means “from” and is used immediately before the callsign of the station transmitting the message. If you hear any response, then find another frequency. Note that some stations will respond with a “Y” for “yes” or a “C” for “confirm”, both meaning “yes this frequency is busy please go away”. Since a lot of operators don’t seem to understand this, or the correct “QRL” in response (“I am busy”), We usually respond “YES” if We am in a QSO and hear someone else call “QRL?”.

If no one responds, then you can call CQ, which is similar to in phone:

CQ CQ CQ CQ DE ZS1AN ZS1AN K

Once again “DE” identifies the callsign of the sending station. The single letter “K” at the end is an invitation to *any* station to reply. Note that the phonetic alphabet is never used in Morse.

Remember only to send as fast as you can comfortably receive at. The station answering your call should come back at roughly the same speed as you send (or slower, if he or she prefers a slower speed). So if you send faster than you can receive you will probably struggle to understand the reply!

Replying to a CQ

As with phone, send the callsign of the station you are calling *first*, and your callsign *second*. For example,

W1XX DE ZS1AN ZS1AN ZS1AN KN

The “KN” at the end with the bar over it means “send the letters K and N together without leaving the normal space between letters”. Since K is dah-di-dah and N is dah-dit, this stands for the symbol dah-di-dah-dah-dit, which is an invitation only to the called station to reply. These symbols made up of two letters run together are known as *procedure symbols*.

The station will proceed to give you a signal report, usually along with his or her name and QTH.

ZS1AN DE W1XX GE OM TNX FER CALL UR RST 439 439 NAME BOB

BOB QTH BOSTON MA BOSTON MA HW? AR ZS1AN DE W1XX KN

As you can see, lots of informal abbreviations are used:

GE - good evening
OM - old man, used to refer to any male operator
TNX - thanks
FER - for, because it is quicker in Morse!
UR - your
RST - RST signal report
HW? - How did you receive this?

The AR symbol with a bar over it is a procedure symbol that means "end of message". It is sounded as di-dah-di-dah-dit. You might reply

W1XX DE ZS1AN R GM BOB TNX FER RPRT RST 569 569 NAME ANDREW

ANDREW QTH CAPE TOWN CAPE TOWN = RUNNING 100W TO 3EL

YAGI = WX FINE 25C 25C = OK? AR W1XX DE ZS1AN KN

Again a few new abbreviations:

R - received everything correctly
RPRT - report
3EL - 3 element
WX - weather
25C - temperature 25 degrees centigrade
OK? - did you receive this OK?

Note that the single "R" sent at the start of the message means "I received everything you send correctly". It is not necessary to spell this out; and conversely, you should not send "R" if you did not receive *everything*. The "=" sign stands for the "break" symbol dah-di-di-di-dah that is usually used to separate thoughts or sentences.

Ending the QSO

You can send the Q-code "QRU?" ("do you have anything further for me?") to indicate politely that you have run out of things to say and would like to end the QSO. Conversely, if a station sends QRU? that is not an invitation to tell him your life story, but an indication that he or she wants to finish the QSO. So it might go like this:

ZS1AN DE W1XX R TNX FER INFO ES NICE QSO QSL SURE VIA BURO =

73 TO U ES URS ES HPE CUAGN = QRU? AR ZS1AN DE W1XX KN

A few more abbreviations:

ES - and (it's shorter and faster in Morse)
URS - "yours" so "U ES URS" means "you and yours"
HPE - hope, or I hope
CUAGN – see you again, so "HPE CUAGN" means "I hope to see you again"

Then we finish with:

W1XX DE ZS1AN R TKS BOB QSL OK VIA BURO 73 ES CUL MY FRIEND

VA W1XX DE ZS1AN TU

"CUL" means "see you later" and the procedure symbol VA means "end of QSO". The "TU" at the end is a final "thank you" and is often followed by two Morse dits – "dit dit" as a final flourish!

Repeater Procedures

Repeaters are used for local FM communication. They allow stations that might not have "line of sight" propagation to each other to still make contact as the repeater will relay the signal between the stations, as long as both have line of sight to the repeater. This is particularly useful for mobile stations with low power and small antennas although it does benefit fixed stations as well.

The repeater is activated ("keyed") when it receives a signal on its input frequency; this signal will then be simultaneously transmitted by the repeater on its output frequency. All the 2 m repeaters in South Africa use a separation of 600 kHz between the input and output frequencies, with the input frequency (the frequency on which the repeater receivers signals) being 600 kHz below the frequency on which it retransmits the signal. When referring to the frequency of a repeater it is standard practice to refer to the repeater *output* frequency. So if someone mentions the "145,750 MHz repeater" they mean a repeater with an output frequency of 145,750 MHz and an input frequency 600 kHz below that at 145,150 MHz.

Overseas repeaters sometime require a tone burst to trigger them. This is not needed in South Africa, so be sure to deactivate your radio's tone burst function if it is fitted with one. However some repeaters do need a CTCSS tone (a continuous sub-audible squelch tone) of 88,5 Hz to activate them, so you may need to set your radio to transmit a CTCSS tone to get into your local repeater.

QSOs on the repeater are much less formal than HF phone QSOs. You don't call CQ on a repeater, for instance – you either call a specific station or just ask if there is anyone listening. For example, We might say:

> *This is ZS1 Alpha November – is there anyone on frequency?*

Note that We have not bothered to use phonetics for "ZS1" – since repeaters are mostly for local use, everyone will know that We have either a ZS1, ZR1 or ZU1 callsign. However We still use phonetics for the "AN" to avoid confusion with, for example, ZS1AM.

If there is a station listening who wants to chat, then he might reply

> *ZS1AN this is ZS1 Bravo. HI there Andrew, nice to hear you again, what have you been up to?*

You do not normally give signal reports when using a repeater, since you do not know what the strength of the other station's signal is to the repeater – you only know the strength of the repeater's output, which is of little interest to the other party. If someone does ask for a report, don't give an RST report but rather tell him or her in plain language that they are "loud and clear" or "some hiss" or "breaking up", whatever the case may be. If you say "full quieting" this means you are receiving a clear signal without any hiss on it.

From there on it is pretty much like a phone call. You should use a minimum of jargon, speak naturally, and remember to give your callsign on every transmission.

There are a number of special points regarding repeater QSOs:

1. The repeaters are a shared resource; they are not there for your private use. So if you want to have a long conversation with one other person then if possible change to a simplex frequency – that is, a frequency on which you communicate directly with the other station, not tying up a repeater. Long "group chats" where anyone can participate are accepted on repeaters and are quite common.

2. Leave a pause of 2-3 seconds after the end of the previous transmission before you start another over. This is to give anyone else who wants to join in the conversation the opportunity to do so.

3. A station that wants to join a conversation or net (network – a number of stations chatting) should wait for a pause between overs, and then just give their callsign, once, on the repeater. The next station to transmit should acknowledge the "breaker" and hand over to him as soon as convenient.

4. If you have an urgent message to pass on a repeater, wait for a pause and then say "break break" and your callsign. The next station to transmit should then hand over to you immediately.

5. Keep your over fairly short, with a maximum of 2-3 minutes. Don't give speeches or sermons over the repeater!

General Points

Remember that the purpose of these procedures is to facilitate clear, intelligible communication even under poor conditions. Do not use unnecessary jargon when plain language will do. For example, the Q code assists with effective communication in Morse code, where it might take too long to spell it out in full. But in phone, it is usually just as quick to use normal language, and more understandable.

Amateur radio is *not* the place to discuss contentious issues such as religion, politics or anything that anyone might regard as indecent. It is fine to have a private QSO with someone whom you know shares your religious beliefs and to discuss these beliefs with them; but it is not alright to "preach" to people you meet on amateur radio. With politics, it is probably best left well alone, even if you know the other person shares your views, as you will never know who else is listening.

Never use insulting, obscene or insulting language, even when you might on the telephone. In amateur radio there is no such thing as a private conversation and all amateurs have an interest in keeping our bands free of abusive and obscene language. If you are heard using unacceptable language even in a "private" conversation with someone who does not mind your language, other amateurs (myself included) will report you to the authorities and ask that your license be revoked.

By law you may not interfere with other QSOs. That includes holding the mike key in to "key over" another transmission. If you think that someone else is hogging the repeater, then by all means point this out politely to him or her. But do not respond by attempting to "key over" their signal, that is a breach of the regulations and if we find you out you will lose your license.

The bands allocated to amateurs are divided into segments for different uses according to the *band plan*. Typically each band will have different segments set aside for CW, digital modes and phone. Some frequencies may be reserved for beacons (which means you should not transmit on these frequencies for any reason) and others may be reserved for particular purposes, such as satellite use or inter-continental DX. Although in most cases it is not a legal

requirement to observe the band plans, courtesy to other operators should be sufficient reason to do so.

Revision Questions

1 The term CQ is used to:
a. Call for a contact with another amateur station.
b. Terminate a conversation.
c. Interrupt a conversation.
d. Make a test transmission.

2 Prior to transmitting a licensed operator should always:
a. Check earthing.
b. Check antennas.
c. Check power supplies.
d. Listen to check whether the frequency is clear.

3 To ensure the calling stations callsign is clearly identified when inviting a contact, the caller should:
a. Repeat his callsign several times.
b. Speak very quickly.
c. Use maximum speech compression.
d. Use the highest frequency.

4 A signal report of 5 9 9 is given when a received signal has:
a. A poor signal strength with a good CW tone.
b. A good signal strength but a poor CW tone.
c. Totally unreadable CW.
d. A perfectly readable, strong and clear tone signal.

5 In the RST code the T is for:
a. Temperature.
b. Tone.
c. Time of transmission.
d. Transmitter type.

6 A readability report of 2 would indicate:
a. Unreadable.
b. Only readable with considerable difficulty.
c. Readable with only slight difficulty.
d. Perfectly readable.

7 The S report in the RST code is obtained from:
a. The power level of the transmitted signal.
b. The speed at which CW is sent.
c. The level of interference on the band.
d. The indication on the receivers S-meter reading.

8 A 59 report is commonly given to stations who:
a. Generate poorly readable signals.
b. Are unreadable.
c. Put in good strong well understood signals.
d. Send poor CW.

9 The term "5 and 9" used to describe a signal, is in which code?
a. Q code.
b. RST code.
c. Morse code.
d. Colour code.

10 The Q code for "standby" is:
a. QRN.
b. QRM.
c. QRS.
d. QRX.

11 QRP means:
a. Close down.
b. Address is.
c. High Power.
d. Low Power.

12 QRT means:
a. Close down.
b. Stand By.
c. Fading.
d. Low Power.

13 Shall I decrease power may be transmitted as:
a. QRT.
b. QSP.
c. QRP.
d. QTR.

14 Will you tell me my exact frequency may be transmitted as:
a. QSL.
b. QRG.
c. QRI.
d. QRU.

15 The use of the Q code is primarily to:

a. Stop unlicensed users understanding transmissions.
b. Save transmitting power.
c. Ensure effective communication.
d. Utilize sidebands.

16 The correct Q Code for "change frequency to" is:
a. QSR.
b. QSX.
c. QSY.
d. QTH.

17 What is the correct Q Code for "what is your location?"
a. QSY.
b. QSP.
c. QRP.
d. QTH.

18 QRM could relate to:
a. I am inundated with static.
b. I am being interfered with by another station.
c. I am going to do a musical transmission.
d. I need more modulation.

19 QRT is defined as:
a. I am going to send now.
b. I am going to stand-by.
c. I intend ending this transmission.
d. Are you going to send now?

20 Which is the correct Q-Code for "shall I stop sending?"
a. QRL.
b. QRK.
c. QRV.
d. QRT.

21 Which is the correct Q-Code for "when will you call me again?"
a. QSD.
b. QSB.
c. QRX.
d. QRH.

22 Which is the correct Q-Code for "are my signals fading?"
a. QSD.
b. QSB.
c. QRN.
d. QRH.

23 Which is the correct Q-code for "Are you ready?"
a. QRL.
b. QRK.
c. QRV.
d. QRG.

24 Which is the correct Q-Code for "can you acknowledge receipt?"
a. QRL.
b. QRK.
c. QRV.
d. QSL.

25 Which is the correct Q-Code for "shall I send more slowly?"
a. QRS.
b. QRK.
c. QRV.
d. QRP.

26 You switch your radio set on and all you hear is a station's call sign in telephony. Do you call:
a. QRA.
b. QRZ.
c. "Who is the station on this frequency"?
d. "Who is calling me"?

27 You are a new amateur and you hear all sorts of phrases being used by other amateurs. Do you?
a. Follow suit and use these expressions.
b. Accept them as correct and acceptable.
c. Add to the vocabulary new words that you make up.
d. Use plain language with normal meanings.

28 Which is the correct phonetic spelling of the word PLUG?
a. Peter Lima Union Golf.
b. Papa Lima Uniform Golf.
c. Pope Lima Uniform Golf.
d. Power Lima Uniform Golf.

29 Which of the following is incorrect usage of the phonetic alphabet?
a. Bravo.
b. Sierra.
c. America.
d. India.

30 Which of the following is the correct phonetic spelling for the word SHIP?
a. Sugar Hotel Item Papa.
b. Santiago Honolulu India Papa.
c. South Hotel India Papa.
d. Sierra Hotel India Papa.

31 Callsigns using phonetics can be given:
a. On every transmission.
b. On the first contact with a station.
c. At the end of each transmission.
d. Regularly.

32 COIL, using the international phonetic alphabet, would be announced as:
a. Charlie, Ocean, Italy, Lima.
b. Charlie, Oscar, India, Lima.
c. Colin, Oscar, Indonesia, London.
d. Colin, Oscar, India, London.

33 Which of the following uses the International Phonetic alphabet?
a. Boston, Uniform, Golf.
b. Bravo, Union, Gold.
c. Berlin, Uncle, Golf.
d. Bravo, Uniform, Golf.

34 Repeaters only normally operate on which mode:
a. AM.
b. FM.
c. SSB.
d. CW.

35 Repeaters are generally operated in the RSA by:
a. A tone burst.
b. A signal on the input frequency.
c. A signal on the output frequency.
d. Remote control.

36 Continuous operation of a repeater by one station is:
a. Desirable.
b. Impossible.
c. Dangerous.
d. Inconsiderate.

37 Satellite frequencies change while monitoring the satellite's signals during its bypass. This is due to the:
a. Height of the satellite.
b. Doppler frequency shift.
c. Drift.
d. The circular orbit shape.

38 Both Azimuth and Elevation refer to:
a. Satellite ground station antenna positions.
b. Mobile communications.
c. Maritime communication.
d. Doppler direction finding.

39 Satellites contain transponders which relay:
a. Only CW signals.
b. Only FM signals.
c. All modes of modulation.
d. Digital signals only.

40 Where should the green wire in an ac linecord be attached in a power supply?
a. To the fuse.
b. To the "hot" side of the power switch.
c. To the chassis.
d. To the meter.

41 What safety feature is provided by a bleeder resistor in a power supply?
a. It improves voltage regulation.
b. It discharges the filter capacitors.
c. It removes shock hazards from the induction coils.
d. It eliminates ground-loop current.

42 For safety in any radio installation it is good practice:
a. To only use plastic piping for earthing.
b. To use unearthed metal piping.
c. Unearth all metal cases.
d. Install a master safety switch known to all in the house.

43 For safety reasons, all exposed metal work in an amateur station should be:
a. Connected to the mains neutral.
b. Free of earth connections.
c. Left completely floating.
d. Connected to a good earth.

44 When wiring up equipment:
a. Any wire available will do.
b. All plastic or insulated wires are suitable.
c. Insulated wires, suitable for the voltages, must be used.
d. Uninsulated wires are suitable.

45 Switches for breaking mains current should be:
a. Single poled and the live leads only broken.
b. Single pole low amperage switches.
c. Double poled and both live and neutral leads broken.
d. Knife switches without covers for easy access.

46 When plugs are used to connect transmitting equipment requiring high current to the mains:
a. Two pin 5 A plugs without an earth pin are suitable.
b. 10 A three pin plugs can be used.
c. Wires can be put directly into the female plug.
d. A 16 A three pin plug should be used.

47 Radio Frequencies are used in micro-wave ovens for cooking purposes. In a radio station care must be taken:
a. To ensure that the power is on by touching RF points with wet fingers to feel or voltage.
b. To Work on RF equipment with the covers off.
c. To adjust antennas whilst full power is applied to the antenna.
d. To screen off all RF source from facial and bodily contact.

48 High capacitance capacitors left on work benches or other available places should be:
a. Passed to another person whose bodily contact can cause a reaction.
b. Should be stored away whilst under load.
c. Should be left lying around with impunity.
d. Should be discharged and stored.

49 When tuning up a transmitter prevent annoying or jamming other users on the band, by tuning, initially:
a. On a harmonic outside the band.
b. Directly into an antenna.
c. Into a dummy load.
d. Directly into a dipole.

50 Amateur Band Plans are formulated and should be observed because:
a. They are mandatory.
b. They are governed by international regulations.
c. They are intended to aid operating and help to avoid congestion.
d. They are there for Novice use.

51 Before making a CQ call on any frequency one should:
a. Send a 1 750 tone burst.
b. Keep giving your call sign repeatedly.
c. Listen on the frequency before and if clear, commence to call.
d. Give your call sign three times.

52 The purpose of a terrestrial repeater is to:
a. Increase satellite coverage.
b. Increase the range of mobile stations.
c. Increase the range of fixed stations.
d. Minimise contacts by pedestrian stations.

53 When calling another station it is accepted practice to:
a. Give your call sign first and then the station being called.

b. Use only your call sign.
c. Give the call sign of the station being called first followed by your own call.
d. Use the call sign of the other station once only.

54 When signing off with another station it is accepted practice at the end of the contact to:
a. Give your call sign first and then the other station.
b. Give the other stations' call sign after your call sign at the end.
c. Don't use the other stations call sign or yours but say over and out.
d. Give the other stations' call sign first and your call sign last.

55 The RST Code translates to:
a. Readability, Signal strength, Tone.
b. Radio, Signal, Time.
c. Readability, Signal strength, Time.
d. Reactivity, Speed, Tone.

56 The RST code "S" is for:
a. Signal Strength of side band signal.
b. Strength of transmitted signal.
c. Safety.
d. Received signal strength.

57 A signal report of "599" means:
a. The signal is of low quality and strength and tone.
b. The signal is of very high quality, strength and tone.
c. The signal is of low strength but good quality.
d. The signal is not easily readable.

58 When calling another station how often should the call sign of the station being called be given?
a. Five times.
b. Three times.
c. Two times.
d. Four times.

59 Once having established contact with another station on a Calling Frequency, it is good practice to:
a. Continue the contact on the same frequency.
b. Move to another frequency and have a QSO.
c. Invite others to join you on the same frequency.
d. Be objectionable to all other stations calling.

60 When two stations are in QSO you should:
a. Butt into the conversation without knowing what they are discussing.
b. Listen first and after finding out the gist of the QSO ask to join and start talking about something else.
c. Butt in and start an argument about another subject.
d. Listen first and if you can contribute to the QSO ask to join and add what you can to stimulate further discussion.

61 Before you come on the air for the first time you should:
a. Know all the procedures used on CB and use them to the full.
b. Use all CB terms even if they do not apply to Amateur Radio.
c. Learn all amateur radio procedures and terms first and only then venture on to the air.

d. Learn all commercial radio terms and use them.

62 When you call CQ for the first time and do not get a reply you should:
a. Move up or down the band and call every few kHz.
b. Call again and again on the same frequency.
c. Change to another band.
d. Listen around the band to see if there are other stations active before calling CQ and call a few times before quitting.

63 The subject matter for any discussion on amateur radio, should include:
a. Politics, religion and sex.
b. Discuss offensive matters.
c. Use indecent language as often as possible.
d. Matters of mutual interest and of a personal or technical nature in a relaxed and dignified manner.

64 When working via a Satellite you should:
a. Use the maximum power permissible.
b. Speak Esperanto.
c. Use sufficient power to maintain reliable communications.
d. Use a speech processor and shout for greater penetration.

65 When using Morse Code initially:
a. Send CQ and your call sign at a very fast speed.
b. Send your CQ and call sign at the same speed that you can receive.
c. Send AK first and then the CQ call.
d. Send AR first and then the CQ call.

66 If a station is calling "CQ Europe" you should:
a. Call him anyway.
b. If he does not answer your ZS call, curse him and accuse him of being anti South African.
c. Wait and see if he gets replies from Europe and if not wait to hear what area he calls next, and so on until he calls CQ Africa.
d. Blow a trumpet or musical instrument to attract his attention.

67 A net is taking place on 2 m. Should you:
a. Call CQ on that frequency during a break in transmissions.
b. Listen for a while and then butt in even if you cannot contribute to the discussion.
c. Wait for a break in transmission, then call in and wait to be called in.
d. Whistle or use a musical instrument to attract attention.

68 When using a repeater on VHF it is good practice to:
a. Use simplex and tell the other stations they are weak and you don't hear them at all.
b. Use maximum power and call until someone answers.
c. Use the duplex mode, and call on the input frequency and listen on the output frequency.
d. Use repeater reverse and hope for the best.

69 When using a repeater you should give:
a. Signal strength reports to other stations.
b. Request a RST signal report on your signal.
c. Give RST signal reports to other stations.
d. Report that you are copying loud and clear.

70 Which of the following is correct, using telegraphy, to make an overseas contact?
a. CQ CQ CQ de ZS1XYZ ZS1XYZ Z1XYZ.
b. CQ DX CQ DX CQ DX de ZS1XYZ ZS1XYZ ZS1XYZ K.
c. CQ DX DX DX de ZS1XYZ AR.
d. CQ DX de ZS1XYZ K.

71 Which one of the following is correct using telephony to make a S.African contact?
a. CQ CQ CQ. This is Zulu Sierra six Zulu Zulu Zulu calling, Zulu Sierra Six Zulu Zulu Zulu calling CQ and standing by.
b. CQ CQ. Zulu Sierra Six Zulu Zulu Zulu standing by.
c. CQ DX CQ DX CQ DX this is Zulu Sierra Six Zulu Zulu Zulu.
d. CQ CQ CQ This is Zulu Sierra Six Zulu Zulu Zulu and Zulu Sierra Six Zulu Zulu Zulu is standing by.

72 Which one of the following is not correct?
a. It is important to speak clearly and not too fast when the other person cannot speak the same language as you.
b. The use of CW abbreviations and Q codes should not be used in telephony contacts.
c. Ham jargon and slang should be used to confuse non amateurs.
d. Avoid the use of "we" when I is meant.

73 When using a repeater it is correct procedure:
a. To pause between overs to permit another station to break in.
b. To transmit immediately after the station in the contact turned it over to you to prevent unwanted stations interrupting your conversation.
c. To pause for a considerable time before replying to the other station in the contact.
d. To monopolise the repeater to prevent others from using it.

74 It is good practice when using a repeater:
a. To use an inefficient antenna.
b. To use a faulty microphone.
c. To use a radio set that overdeviates.
d. To be polite and allow other stations to join into the conversation.

75 When you wish to pass an urgent message over a repeater that is in use you should:
a. Press the microphone switch and shout that you are in a hurry.
b. Whistle continuously to draw attention.
c. Wait until the end of the over, identify yourself, and announce that you have an urgent message.
d. Press the microphone switch and wait until both stations become silent and then take over and pass your traffic.

76 When using a repeater you are told that your signal is breaking up and unreadable. Do you then:
a. Tell the other station that there is nothing wrong with your set.
b. Sign clear until you get into a better position and can access the repeater correctly.
c. Ask someone to relay your unimportant message.
d. Irritate everyone by asking for repeats of messages.

77 When using a repeater one should always:
a. Keep the overs as longs as you feel like.
b. Discuss subjects including politics, sex and religion.
c. Keep the overs short so as to allow other users access.

d. Access the repeater without giving your call sign.

78 When operating on any Amateur Radio band one should:
a. Operate wherever is convenient and unoccupied.
b. Use Lower Sideband in the Upper Sideband portion.
c. Follow the accepted Band Plan for the band being used.
d. Use CW in the phone portion if the band is clear.

79 When operating on High Frequency bands it is good practice after contacting a station initially to:
a. Go straight ahead with the conversation.
b. Exchange signal reports some time during the conversation.
c. Exchange signal reports and ascertain that signals are suitable for a contact before proceeding.
d. Move slightly off frequency to enable the other parties to hear better.

80 When conditions are good and signals are strong operators should:
a. Increase power to the maximum permissible by regulation.
b. Increase power and use full compressor facility.
c. Use only sufficient power to make a good contact.
d. Increase power to the maximum capability of the equipment.

81 Before commencing to transmit, which of the following procedures is not correct?
a. Check that the antenna system is in proper order.
b. Check that there is no undue reflected power on the antenna.
c. Check that the correct frequency is to be used.
d. Assume that the last settings of controls is suitable.

FORMULA SHEET

This formula sheet will be provided to candidates in the Class A examination and may be used to answer any question.

Students therefore do not need to memorise formulae. However the student will need to know which formula to use to solve a particular calculation and will also require the skill to correctly apply the formula. No calculation question will be asked that cannot be answered using the formulae below. In some instances however you may need to use a combination of formulae to solve a problem.

$R_T = R_1 + R_2 + R_3$	$\frac{1}{R_T} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}$	V=IR
$V_{OUT} = V_{IN} \frac{R_2}{R_1 + R_2}$	$P = VI = \frac{V^2}{R} = I^2 R$	$V_{RMS} = \frac{V_{PEAK}}{\sqrt{2}}$
$\frac{1}{C_T} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}$	$C_T = C_1 + C_2 + C_3$	$X_C = \frac{1}{2\pi fC}$
$L_T = L_1 + L_2 + L_3$	$\frac{1}{L_T} = \frac{1}{L_1} + \frac{1}{L_2} + \frac{1}{L_3}$	$X_L = 2\pi fL$
$f = \frac{1}{2\pi\sqrt{LC}}$	$t = \frac{1}{f}$	$\tau = CR$
$Q = \frac{2\pi fL}{R} = \frac{1}{2\pi fCR}$	$Q = \frac{f_C}{f_U - f_L} = \frac{centre\ frequency}{bandwidth}$	$V_S = V_P \frac{N_S}{N_P}$
$R_S = \frac{R_M}{(n-1)}$	$I_S = I_P \frac{N_P}{N_S}$	$R_P = R_L \left(\frac{N_P}{N_S} \right)^2$
$I_C = \beta I_B$	$\text{Gain (loss)} = 10 Log_{10} \frac{\text{power out}}{\text{power in}}$ dB	$I_{RMS} = \frac{I_{PEAK}}{\sqrt{2}}$
$c = 3x10^8$ m/s	$\text{Gain (loss)} = 20 Log_{10} \frac{\text{voltage out}}{\text{voltage in}}$ dB	erp = power x gain (linear)
$V = f\lambda$	$\lambda = \frac{300}{f_{MHz}}$	$\lambda = \frac{c}{f} = ct$

